DOCUMENT RESUME

ED 299 888 HE 021 814

AUTHOR	Adelman, Clifford, Ed.; And Others
TITLE	Performance and Judgment: Essays on Principles and Practice in the Assessment of College Student Learning.
INSTITUTION	Office of Educational Research and Improvement (ED), Washington, DC. Office of Research.
REPORT NO	OR-88-514
PUB DATE	88
NOTE	328p.
AVAILABLE FROM	Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402.
PUB TYPE	Collected Works - General (020) -- Viewpoints (120) -- Information Analyses (070)
EDRS PRICE	MF01/PC14 Plus Postage.
DESCRIPTORS	College Students; Computer Oriented Programs; *Educational Assessment; *Evaluation Methods; General Education; Higher Education; Majors (Students); Motivation; *Outcomes of Education; Program Development; School Effectiveness; Student Characteristics; Student Improvement; Values
IDENTIFIERS	*College Outcomes Assessment; Educational Indicators; Value Added

ABSTRACT

Major technical issues concerning the assessment of student learning in higher education are presented in light of the authors' own knowledge of principles and methods of testing and measurement. This information is intended to help academic administrators and faculty draft a charge to a committee for the design and implementation of an assessment program; to converse intelligently with consultants and faculty members charged with writing the specifications; and to evaluate both the program design and the results of its implementation. Following an introduction by Clifford Adelman, essay topics are as follows: "Designing a College Assessment" (Jason Millman); "Diverse and Subtle Arts: Assessing the Generic Outcomes of Higher Education" (Leonard L. Baird); "Assessment of Basic Skills in Mathematics" (Mark I. Appelbaum); "Issues in Evaluating Measures of Basic Language Skills for Higher Education" (Stephen Dunbar); "Assessing General Education" (John Centra); "Assessment through the Major" (Mark I. Appelbaum); "Assessing Changes in Student Values" (Jerilee Grandy); "Indicators of Motivation in College Students" (Sandra Graham); "Difficulty Levels and the Selection of General Education Subject Examinations" (Clifford Adelman); "Value Added: Using Student Gains as Yardsticks of Learning" (Leonard L. Baird); "Computer-Based Testing: Contributions of New Technology" (Jerilee Grandy); "States of Art in the Science of Writing and Other Performance Assessments" (Stephen Dunbar); "Using the Assessment Center Method To Measure Life Competencies" (William C. Byham); and "Conclusion: Metaphors and Other Guidances" (Clifford Adelman). Appendix A is an annotated bibliography of 79 items (Gary Pike); appendix B contains data on 22 selected assessment instruments (Gary Pike). (SM)

PERFORMANCE AND JUDGEMENT:

Essays on Principles and Practice in the Assessment of College Student Learning

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

U.S. Department of Education
Office of Educational Research and Improvement

Office of Research

BEST COPY AVA

Ordering Information

To obtain information about ordering copies of *Performance and Judgment*, please contact the GPO Order Desk at (202) 783–3238 between the hours of 8:00 am and 4:30 pm, Monday through Friday, or write to the U.S. Superintendent of Documents, Washington, DC 20402.

PERFORMANCE AND JUDGMENT:

Essays on Principles and Practice in the Assessment of College Student Learning

Clifford Adelman
Mark Appelbaum
Leonard Baird
William Byham
John Centra
Stephen Dunbar
Sandra Graham
Jerilee Grandy
Jason Millman

Resource Appendices by Gary Pike

Edited by Clifford Adelman

Office of Research

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402

U.S. Department of Education
William J. Bennett
Secretary

Office of Educational Research and Improvement
Chester E. Finn, Jr.
Assistant Secretary

Office of Research
Sally B. Kilgore
Director

Information Services
Ray Fields
Director

CONTENTS

Acknowledgements		iv
About the Authors		v
Required Reading: Our Introduction	Clifford Adelman	1
Designing a College Assessment	Jason Millman	9
Diverse and Subtle Arts: Assessing the Generic Outcomes of Higher Education	Leonard L. Baird	39
Assessment of Basic Skills in Mathematics	Mark I. Appelbaum	63
Issues in Evaluating Measures of Basic Language Skills for Higher Education	Stephen Dunbar	79
Assessing General Education	John Centra	97
Assessment Through the Major	Mark I. Appelbaum	117
Assessing Changes in Student Values	Jerilee Grandy	139
Indicators of Motivation in College Students	Sandra Graham	163
Difficulty Levels and the Selection of General Education Subject Examinations	Clifford Adelman	187
Value Added: Using Student Gains as Yardsticks of Learning	Leonard L. Baird	205
Computer-Based Testing: Contributions of New Technology	Jerilee Grandy	217
States of Art in the Science of Writing and Other Performance Assessments	Stephen Dunbar	235
Using the Assessment Center Method to Measure Life Competencies	William C. Byham	255
Conclusion: Metaphors and Other Guidances	Clifford Adelman	279
Appendix A: Annotated Bibliography	Gary Pike	295
Appendix B: Data on Selected Assessment Instruments	Gary Pike	313

ACKNOWLEDGEMENTS

The labors of our team of writers, support personnel, and reviewers would not have come to fruition, indeed, would never have taken place were it not for the support and wise counsel of members of the senior staff in the Office of Educational Research and Improvement: Salvatore Corrallo, Director of the Division of Higher Education; Sally Kilgore, Director of the Office of Research; and Bruno Manno, Chief of Staff. Ultimately, though, this project and its allied research efforts owe their being to Assistant Secretary Chester E. Finn, Jr., who pushed and prodded and challenged until this volume was worthy of the trust American higher education places in us to sponsor quality work.

ABOUT THE AUTHORS

Clifford Adelman is a Senior Associate in the Office of Research, U.S. Department of Education. Before joining the Department in 1980, he taught at the City College of New York and Yale University, and was Associate Dean and Assistant Vice President for Academic Affairs at the William Paterson College of New Jersey.

Mark I. Appelbaum is Bowman and Gordon Grey Professor in the Department of Psychology and the L. L. Thurstone Psychometric Laboratory at the University of North Carolina. Since he began his academic career at Chapel Hill in 1967, he has also served as Associate Dean of both the College of Arts and Sciences and the Graduate School.

Leonard L. Baird has over 20 years' experience assessing students and institutions in higher education, first as a researcher at the American College Testing Program and then at the Educational Testing Service. Currently, he is Professor of Educational Policy Studies and Evaluation at the University of Kentucky.

William C. Byham is president of Development Dimensions International of Pittsburgh, Pa., which specializes in training and development, personnel selection and promotion, and productivity issues. Dr. Byham has written 2 books, 17 chapters and dozens of articles and papers describing and evaluating assessment center methodology.

John A. Centra is Professor and Chair of the Higher/Postsecondary Education Program at Syracuse University. He has also been a senior research psychologist at the Educational Testing Service, and, much earlier, an Assistant Professor in Institutional Research at Michigan State University.

Stephen B. Dunbar is Associate Professor of Measurement and Statistics at the University of Iowa, and author of a number of articles on measurement issues in the assessment of job performance. Formerly coordinator of undergraduate composition at the Univ. of Illinois, his current research interests are in psychometric methods and language testing.

Sandra Graham is Associate Professor in the Graduate School of Education at UCLA. Her research and writing have concentrated on the study of achievement motivation among members of minority groups. She is a recipient of an Early Contribution Award in Educational Psychology from the American Psychological Association.

Jerilee Grandy is a Research Scientist at the Educational Testing Service. In her 15 years at ETS, she has directed research for the National Endowment for the Humanities, the National Science Foundation, the American Association of Critical Care Nurses, and the Office of Research of the U.S. Department of Education.

Jason Millman has been a Professor of Educational Research Methodology at Cornell University since 1960, and specializes in the measurement of student achievement and faculty evaluation. He has served as chief editor of the Journal of Educational Measurement and as president of the National Council on Measurement in Education.

Gary Pike is Assistant Director of the Assessment Resource Center at the University of Tennessee, Knoxville. He has been a member of the faculties at the College of William and Mary, Denison University, and Southwest Missouri State University.

Required Reading:
Our Introduction

by Clifford Adelman

This volume is the product of a group of individuals who began their work as occasional acquaintances and finished it as colleagues.

Over a period of 10 months we sought to shape a presentation of the major technical issues concerning the assessment of student learning in higher education in light of our own knowledge of principles and methods of testing and measurement, and in a language that would be accessible to college deans, department chairs, and interested faculty.

This volume is designed to assist those individuals in understanding the psychometric and design considerations involved in different types of student assessment. With neither excess obscurantism nor gross simplicity, the essays contained here should provide enough information for academic administrators and faculty to:

- draft a charge to a committee for the design and implementation of an assessment program;
- converse intelligently with consultants and faculty members charged with writing the specifications; and
- evaluate both the program design and the results of its implementation.

At the same time, this volume is not intended to argue for or against any particular type of assessment program, or to be a polemic on behalf of assessment in general in U.S. higher education. We assume that those who will use this volume as a resource are:

- those in public institutions of higher education in States that have either already mandated or are in process of mandating assessment programs;
- those in public and private institutions whose regional or professional accrediting bodies have recently adopted new standards concerning evidence of student learning;
- those in public and private institutions who seek ways to provide credible evidence of student learning as a normal part of institutional self-study and research; and
- those who seek ways to improve assessment for purposes of student learning and growth.

There is considerable evidence to suggest that many institutions and programs are already affected by external requirements and internal impulses for assessment, and are consequently in need of information and guidance.

The basic "argument" of the diverse essays in this volume is that a combination of the content and purpose of any assessment generates technical requirements to which faculty and academic administrators must respond. They must know which questions to ask and why, how to interpret the answers to those questions, and through this process to grasp the virtues and limitations of the instruments and methods selected. This "argument" stems from standards of ethical and academic responsibility in measurement. It says that if you are going to run an assessment program based on recognized scholarly principles and ethical guidelines, a program that will produce credible and helpful information, here are some important matters that must be considered. You cannot be casual about these issues.

The position of the Office of Educational Research and Improvement (OERI) of the U.S. Department of Education with respect to both this "argument" and the information and opinions provided in these essays should be underscored: the Departmental imprimature implies neither endorsement nor recommendation. OERI's responsibility is to sponsor research and provide information that may help improve American education. We can only commend this volume to its readers as worthy of serious consideration.

Origins of this Volume

This volume is, in part, an outgrowth of the recommendations concerning assessment in Involvement in Learning: Realizing the Potential of American Higher Education. Sponsored and issued by the U.S. Department of Education in 1984, Involvement was the first of the recent spate of reports on the status of higher education, and one that paid considerable attention to the question of how we know what college students learn. This theme was elaborated in a series of regional conferences conducted in 1985 for purposes of discussing the recommendations contained in Involvement, and in the first of what has become a series of annual national conferences on assessment in higher education. These conferences, designed by the American Association for Higher Education and funded initially by OERI and subsequently by the Fund for the Improvement of Postsecondary Education, have drawn large numbers of faculty, administrators, and State higher education officials all seeking to understand what assessment means, why to do it, and how to do it.

As part of the initial effort to disseminate information in support of these activities, OERI published a collection of essays, Assessment in American Higher Education, in 1985. This booklet was essentially a primer. It skimmed the surfaces of organizational and cost issues, the various rationales for assessment, and the range of current practices in post-matriculation assessment. That collection, like other essays on the topic appearing over the past few years, was also polemical. It served a constructive purpose at an appropriate time--awakening many in higher education to both the necessity and constructive possibilities of assessment. But the tough questions and necessary guidances regarding technical aspects of method, instrumentation, and use of assessment information still remained to be addressed in the specific context of higher education.

Indeed, a scanning of the available literature revealed three genres: polemics, descriptions of practice, and technical references. In addition, of course, there are generic literatures on test construction, performance assessment, and legal and ethical standards in test use.

The polemics argue for or against assessment in higher education, often with minimal technical consciousness. The "ayes" seem to include those who, in Kenneth Mortimer's words, would "measure everything that moves." The "nays" seem to be those eager to equate post-matriculation assessment with standardized testing, a convenient equation if one is given to test-bashing or seeks to avoid assessment altogether.

The descriptions of practice are helpful if one reads them as artifacts subject to analysis and critique. The problems acknowledged and addressed in this generally fugitive literature are organizational, not technical; and often where the program described does not wholly succeed as an assessment program, it is alleged to be, in disguise, faculty development.

The technical references consist principally of reviews of individual published tests in the Mental Measurement Yearbook or the manuals and studies of individual instruments presented by their publishers. There is much to be gained from this literature once one has narrowed the field of choice. The reviews provide adequate technical analysis and critique, along with basic "audit" data concerning validity studies, administrative procedures, etc. But this literature is obviously divorced from context and the local purpose of assessment. Furthermore, the domain of this literature is limited to published tests.

Content of this Volume: Inclusions and Exclusions

Given the limitations of these three extant literatures, and given the objective of providing a "hypothetical dean" or (more likely) a faculty task force with sufficient information to engage in the tasks enunciated at the beginning of this introduction, the group responsible for this volume decided, at its initial meeting, to do something else.

We elected to present a series of essays examining psychometric and allied issues in the major target curriculum areas of assessment (basic skills, general education, and the major), in emerging assessment methodologies (performance assessment and computer-interactive testing), and in the assessment of major non-cognitive areas of student growth that are included in the institutional mission statements found in most college catalogues. While we agreed that there was to be no set format or protocol for these essays, they would be preceded by a core essay that set forth the basic principles of assessment design (and to which the other essays would refer), and followed by an essay that reemphasized common themes, picked up the loose strands, and at least acknowledged the unanswered questions.

We imagined that the typical user of the volume would read the core essay (Millman's), and then pick and choose others according to institutional, departmental, or personal interest. We did not assume that every reader would be equally interested in all the topics covered in these essays. There is thus a modicum of purposeful redundancy in the collection that is designed for this type of selective reading.

By adding appendices consisting of annotated bibliography and brief technical audit data on the major instruments mentioned in the course of the essays, we sought to provide, under one cover, a helpful reference work for the target audience. It is not a complete reference work, to be sure, and we did not intend it to be. Readers will not discover in these pages everything they always wanted to know about performance assessment or test bias or construct validity, but they will find enough (along with appropriate references) to explore those topics further. Nor will they gather enough detail on the CLEP test in Organic Chemistry versus the American Chemical Society's test in Organic Chemistry to select one (or neither) for a local assessment, but they will gather enough information to know how to analyze those two examinations in light of both purpose and context of the contemplated assessment.

It is important to distinguish between technical and operational aspects of assessment. The essays in this volume stress the former. The latter, encompassing such issues as costs and other resources, utilization of data and institutional planning, legal and collective bargaining issues and the like, receive little emphasis principally because they are covered better elsewhere (e.g. in the extensive writings of Peter Ewell), but also because they would diffuse the focus of these essays.

Because the focus of these essays is on higher education, this volume should not be used in lieu of either an introductory text on educational testing and measurement or a broad overview of testing and assessment in American society. There are many references we could offer for the former (and some are listed in the bibliography). For the latter, there is probably no better reference than the 1982 report of the Committee on Ability

Testing of the National Academy of Sciences, Ability Testing: Uses, Consequences, and Controversies, edited by Wigdor and Garner. The reader should recognize, though, that the focus of the NAS report is wholly on testing, on testing of abilities (sometimes mis-termed "aptitude") as opposed to achievement, on testing of individuals as opposed to assessment of programs, and on the predictive uses of test data, principally for purposes of selection. While the essays in our volume inevitably cover many of the same concepts and methods as are reviewed in the NAS report, they also cover performance assessments and program and institutional evaluations, and are restricted in coverage to post-matriculation events.

In mapping the domain of this volume, we made five major exclusionary decisions. First, we would focus on near-term student outcomes measures, not long-term behaviors, status, income, or attitudes that may or may not have been influenced by collegiate experiences. There are important reasons that inform this decision, on which we should elaborate. Ours is a practical guide. We well understood that most colleges simply do not have the resources to implement meaningful longitudinal studies, though some do conduct helpful surveys of alumni.

More importantly, there is a well-established distinction in the psychological literature between the proximate and ultimate ends of study, and an equally well-established conclusion that the long-term relationship between the specifics of learning and adult social and vocational behavior is not very clear. In our opinion, most of the questions for which assessment programs can provide answers concern what happens to students between matriculation and graduation, not what happens a decade or two later. Besides, the proximate objectives--and not the ultimate goals--of collegiate study determine the very nature of colleges, the ways in which students and faculty behave in those institutions, and the ways in which those institutions organize their resources, support systems, and governance.

Our second decision was to approach the principal subject of assessment in higher education, the college student, qua college student, not as a candidate for promotion in a corporation or public agency. While it is true (as the essays by Dunbar and Byham remind us) that the working world will eventually assess graduates of colleges and community colleges in terms of particular knowledges, skills, behaviors, and attitudes expected in a given work context, the rules of measurement for occupational competence are not exactly the same as those for college student learning. The principal reason is that a college is a fundamentally different kind of organization with a different mission than a corporation or public agency. Faculty goals--and not employer goals--come first in the assessment of college student learning.

Third, we would not issue pronouncements on the politics of assessment--on who decides to assess and who determines what to assess and when. We assume that whoever is deciding or determining can use the information in this volume to make more intelligent decisions. We do believe, however, that not only faculty, but students as well, should play major roles in the design and execution of assessment programs. While this position is elaborated upon in the concluding essay, it is worth noting here that the very nature of assessment involves students as active participants in understanding their own performance.

Fourth, despite their occasional virtues, we chose not to address proximate unobtrusive measures of student outcomes, for example, job placement rates or rates of acceptance to graduate and professional schools. In our opinion, these measures are too often influenced by factors beyond the control of students, faculty, or the institution itself. Changing labor market conditions, the existence of special national fellowship programs, fad and national mood often contribute far more to what students choose to do following graduation than anything they learned. The elasticity of choice is not as great when the student leaves the institution as it is when the student is in the institution.

Unobtrusive measures, however, should be distinguished from unobtrusive methodologies, and the latter do receive a modicum of attention in the essays dealing with the assessment of values (Grandy), motivation (Graham) and assessment in the major (Appelbaum).

Finally, we decided that this would not be a consumer report on individual tests, though nearly all the essays illustrate their principles with analyses of specific instruments or methods. For most of the working time of this project, we used the phrase "higher education assessment audit" as our password. This was a bureaucratically convenient, but not terribly accurate, phrase to describe what we did. In the world of testing and measurement, an "audit" follows a set of specific guidelines and criteria for documentation to ensure that the development and administration of a particular test meets ethical, legal, and professional standards. It is a checklist covering such topics as prior information for prospective test-takers, procedures for testing individuals with handicaps, and provision of information on test reliability. Most of the information resulting from a test audit can normally be found in the reviews of the Mental Measurement Yearbook, and it seemed foolish to duplicate the function of a major reference work.

As a result of this last exclusion, it may appear that we advocate the psychometrically rigorous development and administration of local examinations and performance assessments, as opposed to the use of "off-the-shelf" tests. Faculty, it is said, are both unhappy with published tests and awed by the task of developing technically sound devices of their own, and will run away from assessment rather than deal with either half of that Hobson's choice. We have assumed, however, that readers of this volume have been placed in a position in which choices are unavoidable. It is not our intention to advocate a particular selection, rather to provide the tools with which such selections can be made.

A Brief Guide to the Essays

Having enumerated the exclusions, let me offer a project director's view of the essays and their writers. These are essays, not chapters, which is to say that having agreed on the overall task and the basic "argument" of the volume, each writer interpreted that task within his/her preferred style of analysis and scholarship.

Some of the essays provide significant detail on specific tests (e.g. Centra on the COMP examination); others use bits of instruments for purposes of illustrating technical principles (e.g. Grandy on values questionnaires). Some read like scholarly articles while others read more like travel guides. Some assume that the reader has previously encountered little of the concepts or terminology of measurement (e.g. Millman), while others will force the reader to stretch (e.g. Dunbar).

While we avoided polemics, these essays do not shy away from contentious topics and positions. Baird, for example, contends that "value-added" is a dubious methodology; Appelbaum challenges the relevance of much current practice in the assessment of basic skills in mathematics; and I take the position that the difficulty of examinations can be addressed without a mass of performance data. As for those topics that often attract zealous advocates, e.g. computer-assisted assessment and assessment centers, the essays in this volume (Grandy and Byham) are as conscious of the technical limitations as they are of the theoretical virtues.

Formats of presentation vary as well. Millman's keynote essay and Baird's piece on the generic outcomes of higher education come with end-notes that are essays in themselves. Graham and Dunbar emphasize theoretical frameworks within which to judge the generation and use of assessment data, and challenge the reader to grasp the framework as much as the particulars.

I regard all these (and other) variances as essential to the character of this volume: there are real voices behind these essays; they are authoritative voices; and they "own" their judgments.

Having introduced the reader to the "what" of this volume, let me introduce the "who." The project group consisted of technical writers, support staff, and reviewers. Of the eight technical writers, six are college professors, though this did not mean that they would have an easy time with their assignments. The two writers who were not resident academics (Byham and Grandy) provided a healthy admixture of perspectives.

For our support work and for the "resource appendices" to this volume, we drew on the Learning Research Center at the University of Tennessee-Knoxville, and specifically on the services of Gary Pike. UT-K is one of the most experienced institutions in the country in post-matriculation assessment, and serves as a clearinghouse for information about higher education assessment under a grant from the Fund for the Improvement of Postsecondary Education.

We also involved the Director of the Learning Resource Center at UT-K, Trudy Banta, as a reviewer. In this capacity, she joined Dwight Lahr, Dean of the Faculty at Dartmouth. Banta and Lahr functioned with Joseph Conaty of OERI and me as a "Board of Perspectives" that initially helped shape the project in terms of the concerns of its target audience and subsequently provided critiques and suggestions to the technical writers concerning the drafts of their essays. In this process, Lahr played the role of the self-effacing dean who feigns ignorance of assessment and then tells you precisely what any dean needs to know in order to design an assessment program. Banta was the scholar of assessment who simultaneously had worked with a very mature statewide assessment program and hence was fluent in the ways in which State assessment policy plays itself out--for better or for worse. Conaty, who joined the Board of Perspectives in midstream, reviewed the drafts as a methodologist. And I assure the reader that my experience as an associate dean involved in assessment programs in a State college has not been clouded by eight years of designing, managing, and conducting research for the Federal Government.

In its final stages of preparation, and in keeping with OERI publication policy, we added three external reviewers: another dean (but from a public institution of moderate selectivity); a director of assessment for a State department of higher education; and a noted scholar and writer on assessment policy and practice. These reviewers came at the volume de novo, were asked to consider it as a whole, and to identify its fatal flaws. That this volume rests heavy in your hands today is evidence that while they may have quarrelled here and there, and while they offered many helpful suggestions concerning everything from the order of the essays to the content of particular references, they found none. They would join me and those who contributed directly to the work in endorsing Jay Millman's understated reflection at our final meeting, "You know, I think we have something pretty good here."

Designing a College Assessment

by Jason Millman

This paper addresses nine questions that should be asked and at least nine decisions that should take place when designing, using, and evaluating college assessment instruments that measure student knowledge, skills, attitudes, and interests.[1] I do not identify all decisions (e.g., assignment of responsibility) here, but emphasize those most directly affecting the technical quality of the enterprise. Subsequent essays in this volume will cover the assessment of specific areas of knowledge, skills and attitudes and will provide examples of instruments and suggestions for their use. Still other essays will identify special issues and new approaches. But this paper will provide the terminology, concepts and framework in which to place the more specific ideas and exemplars offered later in this volume.

This paper discusses each decision, beginning with determining the purpose of the assessment. I place purpose first because the optimal design of an assessment follows the function it serves. In each of the subsequent sections, I introduce the key decision, discuss related concepts and issues, and indicate how the decision should be applied within each of four purposes of assessment: placement, certification, course/program evaluation, and institutional evaluation.

Purposes of a College Assessment

The usual purpose of a college assessment is to make an inference on the basis of students' performance.[2] The inference can be directed to three domains. The educator might want to say something about competence with respect to what is intended to be taught (the curricular domain), about the students' level on some more general ability or trait (the cognitive domain), or about expected performance or behavior in some other situation (a future criterion setting). When inferences are to the curricular domain, they might occur before, during or after instruction. These domains of inference are identified in Figure 1.

The other dimension portrayed in Figure 1 is the subject of the inference: an individual student or a group of students. It is in part because purposes differ both in the domain of inference and in focus on individual or group performance that the assessment effort proceeds differently in each case. The four purposes listed inside the rectangles are the ones emphasized in this essay.

Purpose 1: Placement

A placement decision occurs when a student is assigned to one of two or more educational categories. Examples of placement

Figure 1. Purposes of College Assessments of Student Outcomes

Subject of the Inference	**Domain to Which Inferences Will Be Made**				
	Curricular Domain				
	Before Instruction	During Instruction	After Instruction	Cognitive Domain	Future Criterion Setting
Individual	Placement	Diagnosis	Grading and Promotion	Certification	Vocational Counseling
Group		Course and Program Evaluation			
				Evaluation of the Institution	

decisions include assignment to remedial programs, honors programs, or levels of instruction in such disciplines as foreign language or mathematics.

Two placement situations are worth distinguishing. In a quota situation, the number or proportion of students who can be placed into, or selected for, one or more of the courses or programs is fixed. An honors program that is limited to 60 students is a quota situation. In a quota-free situation, any number of students can be assigned to any given course or program. For example, if it is the case that no student need be assigned to the remedial program if all demonstrate the requisite skills, then a quota-free situation is in effect. In practice, a mixture of both situations is usually present.

Purpose 2: Certification

When an institution certifies a student, it stands behind the claim that the student has the competence implied by the certification award. An academic skills test required to proceed to the junior year and a graduation test are examples of assessment instruments for certification. Some certifications, for example, those enabling a student to obtain a temporary teaching license, are based on assessments more diverse than a single examination.

Assessing for certification has much in common with placement testing, since each involves pass-fail decisions about individuals. In contrast to a placement examination, in a certification assessment inferences are directed to a cognitive domain rather than to a single knowledge domain; the institution and the students have a greater stake in the decision; the institution has the obligation to give the students an opportunity to acquire the competencies being assessed; and the students typically have several opportunities to achieve certification.

Purpose 3: Course and Program Evaluation

When the purpose of assessment is to reach a judgment of merit about the course or program itself, not every student needs to be assessed in the same way or even assessed at all. The course or program, not the student (or the instructor)[3], is the principal focus.

Evaluating a course or program typically includes obtaining evidence of students' performance with respect to the curriculum. It can also include information about students' abilities and opinions, hence the rectangle in Figure 1 extends under the column, "cognitive domain," to suggest this broader range to which inferences will be directed. The dotted section of the rectangle indicates that course or program evaluation that takes place during instruction (presumably to improve instruction through mid-course corrections and refinements) will not be emphasized here, as that function is typically reserved for a single or small group of instructors concerned with a specific course.

Purpose 4: Evaluation of the Institution

Although a frequent reason for assessments is to meet state-mandated accountability requirements, a more immediate reason is to judge how the institution is doing with respect to student learning outcomes. While some may be used in placement and certification decisions concerning individuals, the instruments employed in producing data to assess institutional effectiveness include college and university admissions tests, rising junior examinations, college level skill examinations, general education outcome measures, student attitude and opinion surveys, and exit examinations. Performance on these examinations at two points in time is sometimes assessed to obtain a value-added indicator.

Like the certification function but unlike program evaluation, institutional assessment employs measures that reference primarily a cognitive domain. Like program evaluation but unlike certification, institutional evaluation focuses on groups of students rather than on individual students. The dotted portion of the rectangle in Figure 1 for institutional assessment indicates that discussion of long-term outcomes of higher education is excluded from this volume.

* * *

Can an assessment procedure designed for one purpose work for another? Sometimes it can, but rarely is one instrument or design optimum for several purposes. If too many students are tested, the assessment is no longer efficient. If the wrong skills or too narrow a set of skills are measured, the assessment is no longer valid. The educator who wishes one procedure to suffice for more than one purpose can expect to be frustrated since the responses at the decision points in the design of a college assessment differ for each purpose.

What Content Will be Included in the Instrument?

The content of an instrument consists simply of the questions asked, the statements posed, or the performances elicited. The content of the assessment instrument determines the information that it can generate. For this reason, the decision about what to include is extremely important.

Concepts and Issues

No consensus exists regarding the best way to taxonomize content. A useful distinction can be made between knowing, doing, and believing. The modal assessment methods, correspondingly, are conventional tests (recognition items), performance measures (observing a process and rating a product), and self-reported interest and attitude questions. Controversy most often arises over whether to spend the time and money to assess performance. Performance assessment is practically a requirement in some fields, performing and studio arts being clear cases. But in many other fields, a harder choice exists between measuring what a person knows versus what a person can do.[4]

Content is popularly divided into basic skills, subject matter, general education, and higher-level thinking skills. Because these categories overlap, and because general education is particularly vague, I find it more descriptive to place cognitive measures with respect to content on three different continua. One is lower-to higher-level thinking skills, where the former consists of measures of recall and the latter of measures of critical and analytical thinking, ability to apply principles, and the like. A second continuum is the degree to which students entering at the institution should be expected to know the content or to demonstrate the skill.[5] The third continuum is subject-matter specific to subject-matter general, where the former consists of content specific to a given discipline and the latter of content that crosses disciplines, such as reading comprehension ability.[6]

Keep in mind that the purpose of an assessment instrument is to make an inference based on students' performance. Another scheme for categorizing content is according to how closely the assessment content mirrors the knowledge, skill or trait being inferred. Most achievement tests are direct measures of the desired knowledge and would be classified as low-inference measures. Student enrollment data would be a high-inference measure of course effectiveness, since many factors determine enrollments. The Mosaic Comparisons Test, in which students are asked to identify differences in paired mosaic patterns, is an indirect, hence high-inference indicator of "suitability for certain careers in business."[7]

Other content-related questions that should be asked of a contemplated assessment include: How detailed should the definition of the domain of interest be?[8] How difficult should the knowledge questions be?[9] Should any aspect of that domain receive special emphasis? How broad should the assessment be? (E.g., Should untaught subject matter be included? Should measures of unintended outcomes be developed?) The reason to ask such questions is that if one is too narrow in coverage, what one learns from an assessment enterprise is limited.

Let us examine the form these content issues and questions take under each of the four purposes of assessment. That is, given a specific purpose for an assessment, what special considerations concerning content issues should be considered?

Purpose 1: Placement. Two views about the appropriate information for a placement decision dominate practice. One emphasizes previous courses taken and credits earned; the other emphasizes knowledge and skills that can be displayed. I favor the latter, and recommend that placement decisions be based on the tested level of knowledge and skills of the student, not on high-inference proxy measures.

In the quota-free situation, the domain of content should be quite narrow and focus upon that body of knowledge and skills most apt to differentiate among students at the borderline between placement categories. It is sometimes helpful to consider the examination questions answered correctly by students who barely qualify for the more advanced course or program and incorrectly by students who do not qualify. Alternatively, one can consider those students who barely miss qualifying for the more advanced course or program and analyze their responses. In short, the placement examination in the quota-free situation should concentrate at that level of functioning near the borderline between categories.

In the quota situation, the function of assessment is to identify the highest "X" number of qualifying students, where X is a value determined before the assessment begins. If X is a small fraction of the students, then the questions or tasks should be somewhat difficult so that it will be possible to differentiate reliably among the able students. Similarly, if X is a large fraction of the students, then the questions or tasks should be somewhat easy.

Purpose 2: Certification. Regardless of what competencies are being certified, be they academic skills, thinking abilities, or knowing and being able to do what an educated person knows and can do, the competencies should be clearly defined. Students have a right to know what is required of them--not the specific examination questions, but the domain of coverage. Further, clear definition of the content will assist the institution in designing instruction to fulfill its obligation to provide students with the opportunity to acquire the competencies. And the competencies measured by the assessment instruments ought to be worth acquiring.

In determining the domain of coverage, faculty must be sensitive to the diversity of programs in their institution. Knowledge and skills required of all students seeking the same certification, regardless of their differing educational goals, should not be chosen lightly.

Purpose 3: Course and Program Evaluation. A broad range of content can, and probably should, be covered. For example, it is not necessary to limit the measurement of student achievement to a narrow interpretation of subject matter taught. Questions can be posed differently than presented in the textbook or class. Questions can probe whether students are able to transfer what they have learned to situations not covered in the course materials. Unanticipated outcomes or side effects should also be probed. What misconceptions emerge? What new attitudes have been developed? What effects have the course or program had on the students' general abilities? How, if at all, has the course or program changed students' feelings about themselves, about the institution, and about issues related to the subject matter?

The evaluation designers would do well to begin by asking a small group of students and interested others to share their perceptions of the course or program. These perceptions form the grounds for some of the items in the assessment instruments. College students are an excellent source of information about a course or program, as they are first-hand witnesses to the instruction, facilities and materials for an extended period of time. Nevertheless, a more thorough evaluation would include other sources of information than that obtained from tests, surveys, and academic products.[10]

Purpose 4: Evaluation of the Institution. One focus of the debate over the evaluation of the academic effectiveness of colleges and universities is on the content of assessment measures. Should they

> emphasize the acquisition of facts and the mastery of simple skills....[or] how clearly students think about issues of social justice, how deeply they can appreciate a painting or a literary text, how much they have gained in intellectual curiosity, how far they have come in understanding their own capacities and limitations?[11]

The choice of content for the evaluation of the institution says much about the institution's view of its educational mission.

It is not enough to provide a general label for content. The domain must be clearly specified not only so valid measures can be constructed to reflect the desired outcomes, but to guide curricula and instruction. Students typically do not grow into renaissance people, lustful for learning and sparkling in curiosity, without assistance. Promoting growth requires a clear sense of direction.

It may be tempting to let a convenient, commercially-available instrument or an examination now in use at the institution, such as a general education examination, serve as the principal instrument to evaluate the institution. But educators need to ask whether these devices both capture the richness of the educational outcomes desired for students and yield data that can inform the evolution of academic programs.

Will the Instruments Be Developed Locally or Obtained from Another Source?

College administrators and faculty have a choice of whether or not to construct their own assessment instruments. If the decision is to look elsewhere for an instrument, several options are available. One is to purchase an off-the-shelf instrument from a commercial publisher. A similar option is to secure permission to use an instrument developed by a state agency or another college. A third option is to engage the services of a test-development firm or consultant, either inside or outside the institution, to build an instrument to the institution's specifications. Still another option is to work with other institutions to build, jointly, the desired instrument.

Concepts and Issues

Availability, quality, cost, and sense of ownership are four factors that influence the decision whether to develop one's own instrument or obtain it from another source.

Availability means finding an existing instrument that matches the content coverage desired by the institution. Existing instruments are identified in several sources. A bibliography of assessment instruments appropriate for college assessments is available from the University of Tennessee.[12] A more extensive list of tests as well as test critiques are available both in hard copy and on-line computer from the Buros Institute.[13] The Test Corporation of America also offers lists of tests and test critiques.[14] There is even a bibliography of test bibliographies.[15] Even if the decision is to develop the instrument locally, evaluating existing instruments can broaden one's perspective of a domain and how to measure it. For a modest fee, most companies will send prospective users a specimen set that contains a copy of the test together with related material. Some tests are kept secure, however, thus denying educators an opportunity to compare the items on the test against the qualities the institution wants to measure before purchasing rights to the instrument.

A second consideration is quality. Criteria of quality include validity, freedom from bias, and reliability, all of which are discussed later in this essay. A match between the content of the instrument and the content desired is extremely important. Educators should not assume that because an instrument is published, it is of high quality. On the other hand, local educators often have neither the time nor the expertise to write, edit and try out an instrument to the degree commercial publishers do.

An important consideration is the relative cost of purchasing existing instruments versus developing one's own. Exceptions exist, but if the time of faculty/staff and other opportunity costs are factored in, using existing instruments is less expensive. Instrument development is a demanding, time-consuming activity. Noting the approach and items used in existing instruments, however, can make a local development effort more efficient.

Developing one's own instrument, while sometimes difficult, is nonetheless attractive, particularly if resistance to the assessment is anticipated. A sense of ownership and acceptance of the assessment is more likely for a home-grown than for an imported product. The process of creating an examination can have a greater effect on an institution than the examination itself.

Two other factors sometimes considered in this decision are credibility and availability of norms. Some publics consider externally constructed instruments more credible than locally developed ones. The credibility (but not the validity for the institution) rises if the instrument or the publisher is well known. Availability of results from comparison groups (i.e., norms) is considered by some to be a strong reason for purchasing an existing instrument. This advantage may be illusory because the norm groups are usually ill-defined, thus not interpretable, and because normative information is not particularly valuable for most assessment purposes.

If the instrument is to be used for placement or certification decisions, it must be related to instruction, and thus faculty involvement in the decision is important. For these assessment purposes, the decision to purchase a test or construct one's own should depend heavily on the availability of instruments that match the instruction. Fortunately, a number of examinations in basic skills and college subjects are available for purchase. On the other hand, they are also easier to construct locally than instruments that measure harder-to-define domains such as critical thinking, creativity, or aesthetic sensibility.

For course and program evaluation, locally produced instruments will likely be needed to attend to specific concerns about the course or program being evaluated. Such instruments might supplement one or more commercially available ones. For course, program or institutional evaluation, wide faculty and student participation in either constructing or selecting an instrument is important. Since any evaluation (especially an evaluation of the value of the college or university experience) benefits from multiple measurements, it may be that a combination of commercially available and locally developed instruments will work best.

How Will the Data Be Collected?

The general question is, Who will be administered what instruments or questions and how? The "who" and "what" part of the question leads to a consideration of sampling. The "how" part includes questions of student motivation and preparation and test administration.

Concepts and Issues

Sampling. One part of the sampling question is simply a concern about who should be included in the assessment. Sometimes the appropriate group of students is obvious; other times it is not so clear. Will part-time and transfer students be included? Should any subgroups of students be treated differently? The decision about who is to participate will determine the student population to which the results can be generalized. To increase the interpretability of the results, assessment specialists recommend that the rules for inclusion and exclusion be clear and objective.

Another aspect of the sampling question is how many students (in the specified population or group of students) should be included? One answer is to include all the students, perhaps because it is thought to be too much trouble to do otherwise or perhaps because results based on a sample are less credible in the eyes of some people. A different answer is to recognize that random or representative samples can be cost efficient and that an appropriate sample size is the number of students that will produce the degree of precision desired by the user. Random samples of 100 students can yield estimates with a margin of error of 10 percent or less. Quadruple the sample size and the margin of error is cut in half, so that if n=400, the margin of error is 5 percent or less. In reverse, reducing the sample to one-fourth its original size will only double the margin of error. If the margin of error is small to begin with, the loss in precision of estimates by reducing the sample size can be tolerated. The decision to sample only some students saves money at the expense of precision--although precision may be only marginally affected if the sample is large and representative.

An important (though often neglected) aspect of the sampling question is what instruments or questions should be sampled? It may not be necessary for everyone to be exposed to the same assessment device. For some purposes, some students can be administered part of the assessment instruments and other students a different part. This scheme is called "matrix sampling" because a matrix of students and assessment items are sampled. For example, a 120-item survey could be divided into four forms of 30 items each. One-fourth of the students in the sample might be asked to answer each of the forms. Matrix sampling has the advantage of covering the domains of inference more broadly than possible if every student responds to the same, more restricted set of questions. The decision to use matrix sampling increases the scope of the assessment at the expense of precision of any one estimate. In the above example, the scope is increased four-fold and the margin of error associated with any one item is doubled.

Student Motivation and Preparation. Identifying the students is one activity; convincing them to submit to an assessment is another. And motivating them to do their best is yet another. Success in these activities is important if the assessment is to be valid. Experience in the context of certification examinations has clearly demonstrated that adults perform markedly poorer when little is at stake than when an important decision depends upon their performance.

One educator in charge of a college assessment program said to me, "If I had to do it over, I'd have started with the students rather than with the faculty." The suggestion is not only to involve the faculty, but the students as well. Having students and faculty help design the assessment, construct or select the instruments, and even take the tests and surveys is one way to increase participation and motivation. Using tasks that make students think, that offer feedback, or that promise changes is another way. Harvard University had excellent participation from its students in a survey when they were given immediate feedback about their responses and, a short time later, comparative information about other students.[16] Another option is paying students when participation in the assessment is of no obvious benefit to them, but this strategy will likely work only for "one-shot" ad hoc evaluations.

Optimum performance on achievement and aptitude measures is clearly a goal of an assessment program. Valid assessment of student knowledge and skills requires not only motivated students, but students who are appropriately prepared for the assessment tasks. For most assessment applications, students should be informed ahead of time about the general areas being measured (not the specific test questions) and given practice with unusual question formats. If, after a little instruction and practice, students can greatly improve their performance on a particular assessment task, then assessing this enhanced performance is usually more meaningful than assessing their initial level of performance.

Instrument Administration and Security. Assessment will be used to make comparisons. Results will be compared to standards or to the performance of different groups or at different times. For these comparisons to be valid, standard administration practices have to be followed. If more time is given one group than another, if more help is offered one time than another, or if testing conditions on one occasion are noticeably different than on another, the comparative data are contaminated.

Test compromise (i.e., cheating) also invalidates assessment results. Especially in high-stake situations, students' methods of gaining illegitimate help on an assessment exercise are often quite sophisticated. At one extreme, some test sponsors never use a particular instrument more than once, believing that it is impossible to secure the questions. At the least, instruments intended for multiple administrations should be carefully inventoried and kept under surveillance. Physical separation of students at the assessment site and use of multiple assessment forms lessen the risk of test compromise.

Special populations, such as linguistic minorities and the physically handicapped, require special attention. There are published professional standards for testing such groups,[17] and the concluding essay to this volume offers additional guidance.

These data collection issues require different emphases under each of the previously described four purposes of assessment:

Purposes 1 and 2: Placement and Certification. In any given context, the identity of the students who should be administered a placement or certification examination will be obvious. No sampling of students is involved, since all students for whom a placement or certification decision is required should take the examination.

It is important that students do their best so that the decision will be based on knowledge of their true abilities. Motivation is important. These considerations suggest the advisability of distributing advance information about the content and importance of the examination. In particular, topics that can be quickly learned or reviewed should be brought to the attention of students.

Examination administration conditions should be constant, so that the cutscore (passing score) value has constant meaning across administration sites. Adaptive testing approaches (see Grandy's essay on "Computer-Based Testing," below) in which students are administered items close in difficulty to their own level of functioning, have no advantage in the typical quota-free, two-category placement setting or the typical pass/no pass certification situation. Especially since placement and certification decisions can be important to students, guarding against prior circulation of the examination and other instances of test compromise is warranted.

When the purpose of assessment is certification, students should have more than one opportunity to demonstrate competency. The individual stakes are too great for the decision to rest on a single administration of a single, less-than-perfect instrument. Providing students an opportunity to take the examination a year or even more before the deadline date is not unusual.

Purposes 3 and 4: Course, Program, and Institutional Evaluation. In these applications, as previously noted, the performance of individual students is not being compared and the content or domain coverage is broad. Matrix sampling is optimum under these conditions. Allocating different assessment measures to different students permits the evaluator to learn much about the course or program while limiting the burden on any one student.

A possible exception occurs when results are to be disaggregated into units (i.e., courses or departments) having relatively few students. That is, if precise information is wanted about a course with a small enrollment (e.g., 35), essentially all students should participate. Matrix sampling will work well, however, in a department with 300 majors.

A special student sampling concern arises when the evaluation follows the value-added model. In the value-added approach to assessment, effectiveness is measured by the difference in performance of beginning and graduating students. The strongest design compares the results on a constant set of instruments administered to the same group of students when they entered and left the program or institution. Often, however, it is not feasible to assess the same students at two different points in time. A common but flawed approach is to compare unselected groups of beginning and graduating students. These groups are not comparable. Even in the absence of any program or institutional effect, graduating students could be expected to be more able on the average than entering students (a group including many who will drop out). Another common approach is to use ACT or SAT admissions test results to predict what graduating students would have scored on the assessment instrument as beginning students. This estimate is taken as the base from which the value added is computed. In my view, a better approach is to construct a sample of entering students who match the graduating group as closely as possible on indicators of ability or predictors of success, such as admission test scores, high school grades, college major, gender, race, and even initial grades in the college.

Because little is at stake for individual students in a course, program, or institutional evaluation, motivating them to participate and do well can be a particular problem, while providing practice opportunities and assuring test security are likely to be of relatively little concern. Efforts are clearly called for to raise the level of seriousness with which the assessment exercises are taken.

Following rather rigid, formal administration procedures is advisable if the results for different groups will be compared or if the assessment will be replicated in the future. If, however, the purpose of the assessment is open-ended, designed more to generate ideas than to confirm hypotheses, a more informal administration of the instruments is reasonable.

What Additional Instrument-Development Efforts Will Be Needed?

Three activities that are desirable in certain assessment contexts are selecting or constructing additional forms of the instrument, equating forms, and establishing passing scores. The associated questions are: Is more than one form of the instrument needed? If so, Should the forms be equated? and, if so, How? Is a cutscore required? and if so, How will it be set?

Concepts and Issues

Multiple Forms. Multiple forms of an assessment instrument are particularly desirable if any of three conditions exist. The first is when the instrument cannot be kept secure, a condition that exists when a great deal hinges on the outcome. The second is when the goal of the assessment is to measure change and the questions are such that previous exposure to the same items (rather than improvement in knowledge, skills or ability) will lead to improved performance. A third condition occurs when the content of an instrument is no longer current or is judged less representative of a domain due to changes in the curriculum.

Equating. Two instruments are equated if, on the average, it doesn't matter which one is administered. No two instruments can be exactly the same in difficulty, but it is possible to adjust scoring so that, on average, students will not receive higher or lower scores on one form compared to another. If performances (of the same or of different students) on two or more forms of an instrument are to be compared, they should be equated. Otherwise, these differences might be due to the instrument rather than to the underlying ability.[18]

Cutscores. Cutscores are needed if decisions about students depend upon their performance. An exception occurs if a decision is based on several factors and no minimal level of performance on the instrument is required. Many methods of setting a cutscore have been suggested.[19] The most favored are those methods that depend upon judgments by experts who are informed by data about student performance. Although the standard-setting process need not be arbitrary, it does require judgment.

Because no measure is infallible, errors will occur regardless of where the cutscore is placed or how it is established. Students who deserve to pass may be failed (false negatives) and students who deserve to fail may be passed (false positives). Educators in the position of having to make pass-no pass (or select-not select) decisions about students, and who believe that in their situation it is more important to avoid false negatives than false positives, will want a low cutscore so that anyone who fails clearly does not possess the required level of proficiency. On the other hand, educators who believe that in their situation it is more important not to pass anyone who might not possess the required level of proficiency, will establish a high cutscore. Although errors of classification cannot be eliminated, they can be reduced with more valid and reliable instruments, and the ratio of false negative or false positive errors can be controlled by where the cutscore is placed.

It is almost always the case that differential passing rates will result for identifiable groups, such as black and white or male and female. The group with the lower passing rate is said to be adversely impacted by the test. The extent of the adverse impact of an examination depends heavily on the cutscore. No adverse impact would result if the cutscore is so low that everyone is placed in the more advanced course or program or meets the certification requirement, or if the cutscore is so high that no one passes. On the other hand, if the cutscore corresponds to a score midway between the averages of the two groups, adverse impact would likely be very high.

How do these three issues play out under each of the four purposes of assessment?

Purposes 1 and 2: Placement and Certification. The need for multiple forms typically differs for these two assessment purposes. Placement examinations are often given only once to any individual, and often the educator can maintain security of the instrument. This is especially true if the examinees have recently been admitted to the college. Thus, one form of the test may be sufficient and, since only one form is used, equating is not necessary. If the assessments result in placement in remedial courses for which post-tests are used, then obviously more than one form of the tests should be developed and equated.

More than one form of the assessment instrument is strongly suggested in the certification context. Since each student may be given more than one chance to pass the examination, multiple forms mean that the student can be administered different questions on each testing occasion, thus providing some assurance that the test measures the competency intended rather than the ability to learn the answers to one set of questions. Also, security is increased with multiple forms. These forms should be equated in difficulty so that a student's chances of achieving certification do not depend upon which form was administered. Equating test forms thereby ensures a common passing standard.

Cutscores are required for placement and certification purposes because decisions rest on the examination results. The location of the cutscore affects the adverse impact of the assessment, the passing percentage, and the relative mix of false positives and false negatives. Cutscores should therefore be set thoughtfully. Once set and used, instructors can be asked how many of their students, in their judgment, were incorrectly placed or certified. For example, if a much larger than usual number of students in an advanced course or program are so identified, then that would be evidence that the cutscore on the placement examination might be too low. Similarly, if many more students than usual are judged misplaced in a less advanced course or program, that would be evidence that the cutscore might be lowered.

The relative values placed on avoiding false negatives and false positives will help determine the cutscore. If one type of misclassification error is considered more serious than another, the cutscore should be raised or lowered accordingly. In the certification context, the institution has some stake in avoiding false positives, and students may have more than one opportunity to take the examination. For these reasons, the cutscore on a certification examination might be set a shade higher than otherwise.

One variation on the sharp cutscore approach, in which those who score above it are treated differently than those who score below it, is to establish an uncertainty band on either side of the cutscore. Falling in the uncertainty band could trigger the gathering of additional assessment data before a decision is made. In the placement situation, students who fall in that band can be given a choice of course or program.

Purpose 3 and 4: Course, Program, and Institutional Evaluation. Test security is relatively less important for these assessment uses, so only one form of the assessment instruments is needed. Even if a value-added assessment is conducted, it is likely that a sufficient time period will elapse between assessments so that administering the same form of the instrument will be of little consequence. If only one form is employed, equating procedures are not applicable. Finally, cutscores are not required because no decisions about individuals are involved.

How Will Bias Be Detected and Minimized?

Instrument bias is a major concern of many groups in our society, and emotions and preconceptions are strongly rooted. Bias is a source of invalidity. It is important to be clear about what is meant by bias and how to identify and eliminate it from our assessments.

Concepts and Issues

In a rational discussion of bias, three distinctions are helpful to keep in mind. The first distinction is between instrument bias and adverse impact. One racial, gender, or other group may be adversely affected by the assessment. It may score appreciably lower, and thereby be denied opportunities given to those who score higher. Quite often black and Hispanic students score much lower on achievement and aptitude tests than do white students, and for some tasks, women score noticeably lower than men. As a result, these groups may receive fewer scholarships, be placed more often in remedial courses, and the like. Such impact, however, does not mean the instrument is biased. It could be that the adversely impacted groups truly do achieve less in the terms measured by the instrument. Years of educational deficit or other social differences between the groups may have taken its toll on the present knowledge and skills of its members. Because a thermometer records different temperatures in two rooms of a house does not mean the thermometer is biased.

The second distinction is between instrument bias and unfair use. An instrument may be a near-perfect measure of what it is supposed to measure, but nevertheless be used inappropriately. As Appelbaum notes in his essay on basic skills assessment in mathematics in this volume, using the SAT/Quantitative (a perfectly valid measure of learned abilities) to place students into remedial courses is inappropriate, in part, because it would be unfair to students whose general quantitative reasoning abilities are low, but who nevertheless had previously learned the specific content of such courses. Instrument bias is assessed by considering the evidence for the validity of the inferences or uses to which it is put, or by judging the bias of the items that compose the instrument.

Item bias versus instrument bias is the third distinction. Most of the techniques for judging bias are attempts to identify biased items. If the items are found to be unbiased, then the instrument as a whole is considered unbiased. Two broad categories of techniques for detecting item bias are in use.[20] The first category consists of methods by which appropriate groups of raters consider whether the content of the items reflect bias. In addition to judging whether all groups have equal familiarity or experience with the particular examples and language included in the items, the raters often are asked to identify items with stereotypical or offensive content. It is advisable to include among the raters representatives of black, Hispanic, female or any other group likely to be adversely impacted, particularly if the purpose of assessment is certification.

Statistical methods of detecting bias in items that have right and wrong answers is the other category of techniques for detecting item bias. In the methods in this category, the responses of students who are from different gender, racial or cultural groups and who are considered equal in ability are compared. In one method, for example, students receiving the same score on the entire examination are assumed to be equal in ability. If the members of one group consistently miss an item proportionately more often than the members of a second group having the same total examination score, the item is considered biased. For results to be reliable, statistical techniques of bias detection require 200 or more students in each of the groups being compared (if necessary, this number can be achieved by aggregating data from previous years, provided that the same instrument was used). When both categories of methods are applied to the same instrument, they often identify different items as biased. When the purpose of the assessment is certification, the use of statistical indicators of bias is feasible since relatively large numbers of students are involved in the assessment. But it is probably best to think of the statistical results as a supplement to, rather than a substitute for, the judgments of educators.

In the cases of course, program, and institutional evaluation, instrument bias is less of a concern because the evaluations have little immediate effect on students. Bias in the design and implementation of a course or program itself is more apt to be a concern, and data that address this concern might well be built into the evaluation plan.

It may be that the results of an evaluation of the institution will be used to judge the institution unfairly, which happens when the capabilities of entering students are not considered in judging student outcomes. It may also be that some educational segment of the institution will be adversely affected by the evaluation because the educational goals of that segment are not sufficiently represented in the student outcome measures being employed. This possibility underscores the desirability of having a broad representation of the institution serve on an assessment development committee.

How Will the Validity of the Assessment Instruments Be Determined?

The most important characteristic of a college assessment is the correctness--that is, validity--of the descriptions and decisions that emerge from it. Correspondingly important are decisions about how faculty will determine the validity of their assessment instruments and methods.[21]

Concepts and Issues

Technically, it is the inference from the instrument and not the instrument itself that is valid. For example, a test may measure basic skills very well but not measure general education knowledge well. Thus, the inference about basic skills is valid; the inference about general education knowledge is not valid.

The first step in validation is to determine exactly what inferences are desired--what curricular domain, cognitive ability, or criterion behavior is targeted (see Figure 1). The remaining steps consist of gathering evidence that is credible to others and that addresses the accuracy of the inference. Establishing validity is like performing research. The hypothesis is that a given inference is correct, and the effort is to marshal evidence in support of or against that hypothesis.

The evidence that will be credible and relevant depends upon the inference. For example, when the inference is to a curricular domain, properly collected judgments from respected and knowledgeable individuals about the appropriateness of the assessment instrument's content can be persuasive. To take another example, if the results of a placement test are used to place individuals into one of two levels of Spanish, then the validity of the inferences would be supported by evidence that the individuals in the first course were appropriately challenged and those in the second course performed well.

Evidence that will be credible and relevant also depends upon the concerns that have been expressed about the assessment. For example, if the instrument is viewed by some to contain irrelevant content, or to have been administered inappropriately, or to favor one group over another, then information that addresses these concerns should be gathered. Test and item bias are sources of invalidity.

Under each of the four purposes of college assessments, different kinds of validity evidence are important:

Purpose 1: Placement. Two types of validity evidence are particularly relevant for placement examinations, namely, faculty judgments and student performance. Faculty knowledgeable about the subject can examine the congruence between the items on the examination and the skills and knowledge needed to perform in the more advanced course or program. They might be asked the specific question, Do the items measure the prerequisite knowledge and skills? Instructors in the course or program could be asked whether students in the more advanced course or program are sufficiently prepared or whether students in the less advanced course or program appropriately placed?

As for student performance, a high correlation between actual grades in the more advanced course and placement test scores would provide additional evidence for the validity of the placement decisions.

Purpose 2: Certification. Two types of evidence are particularly appropriate for competency examinations. One, content validity, would be revealed by the judgments of a heterogeneous set of faculty that the questions on the examination both match the abilities defined by the domain and are a representative sample of them. A clear definition of the domain is helpful in this task. For example, if a test claims to measure a learned ability such as "solving problems,"[22] then the judgment task is greatly assisted if the types of problems a person with the ability is expected to solve are identified in detail.

A second type of evidence is curricular or instructional validity, i.e. documentation that students had an opportunity to acquire the skills measured by the assessment instrument. Note that the issue is not whether the students actually have acquired the skills, but whether courses or materials were available from which the skills could have been acquired. The courts have held that at the high school level at least, curricular or instructional validity is required for tests used as a requirement for graduation.[23]

College educators should avoid the trap of claiming that specific competencies or knowledge are required for success after college. Success depends upon many factors; academic and related skills is but one configuration of factors. For that reason, finding a high correlation between performance on a certification instrument and future success is very unlikely.

Purpose 3: Course and Program Evaluation. In these applications, we want to know not only whether the assessment instruments are valid, but also whether the evaluation itself is valid. In the case of instruments, we can check whether our inferences about students' subject matter knowledge or attitudes and opinions are correct. A number of techniques have been developed to assist in these validation tasks (see end note 21). For example, if all the items on an assessment measure the same attitude, students who answer an item in a positive direction should be expected to answer the other items in the same direction.

Student response data, by themselves, are insufficient to answer the second question, How valid are the judgments of merit of the course or program? A sound evaluation of a course or program requires additional survey and analysis data that attend to such questions as the need, intrinsic value, relative value, use, costs (direct and opportunity) and future potential for the course or program.[24]

Purpose 4: Evaluation of the Institution. In this application, the validity questions are whether an instrument accurately measures the intended ability or attitude and whether a valid evaluation of the institution was conducted. The first asks if the assessment device measures what it purports to measure.

Judgment of the congruence of the items to the definition of the domain helps. Techniques are available to assess whether the student responses to the instrument follow patterns that are expected if the inferences from the instrument were valid (see end note 21).

The second concern, evaluation validity, asks whether the instruments measure a reasonable range of student outcomes in which the institution should be interested. Judgments of a cross-section of faculty and others are useful in assessing the validity of the evaluation.

Often only a single instrument is administered during the evaluation. In such cases, the inferences from the instrument may be judged more or less valid, but the evaluation itself may be found to be seriously limited in its ability to determine how the institution is doing.

How Will the Reliability of the Assessment Instruments or Methods Be Determined?

Reliability refers to consistency. Valid inferences can be enhanced by reliable measurement. Results which fluctuate from instance to instance cannot be relied upon. Evaluating the reliability of assessment instruments is standard practice in the testing industry, and the task is just as important when assessment methods other than paper-and-pencil tests are at issue.

Concepts and Issues

Three kinds of reliability are of varying importance depending upon the purpose and design of the assessment.[25] The first is the precision of the results. How much fluctuation can be expected in the assessment results (individual scores, group means, measures of differences or change) if the assessment were redone either at a different time or with a different, but similar set of questions? The expected fluctuation is expressed by a statistic called the standard error of measurement.

The second kind of reliability refers to the consistency of decisions. What proportion of the decisions (e.g., passed the test, placed in a remedial program, etc.) would be the same if the assessment were redone, presumably with different, but equivalent instruments? The consistency is often expressed as a simple proportion. A decision consistency of .80 would be interpreted to mean that 80 percent of the students would achieve the same classification decision on the two assessments.

The third notion of reliability refers to the consistency of the raters. How well do the judges who grade or score the assessment agree? Rater or scorer reliability is particularly important when subjective measures of student performance are employed. Experience has shown that when essays or other educational products are graded or when performance is observed and rated, the score values assigned are likely to differ unless standards are agreed upon ahead of time by the evaluators. In assessing rater reliability, the scores should be assigned independently.

The validity of the findings of a college assessment are seriously threatened if its instruments are unreliable. Unreliability means that, depending upon which questions happen to be asked or which raters happen to be judging the student's performance, the scores, decisions, or ratings would be different.

Three facts about reliability are important to keep in mind. The first is that reliability is heavily dependent upon the length of the assessment instrument. If the assessment is found to have unsatisfactory reliability, the condition can be ameliorated by including additional questions similar to those on the instrument or by including additional, but equally trained judges.

Second, measurements about individual students are less reliable than measurements about groups of students. Longer assessments are required when the results affect and will be communicated to individual students than when group averages will be reported. For example, the Academic Profile referenced in end note 6 consists of three, one-hour forms. For individual assessment, the publisher recommends that the forms be combined and administered to each student. For group assessment, the publisher recommends a matrix sampling design in which the forms are randomly distributed within each group such that each student responds to only one form.

The third important fact about reliability is that assessments are less reliable when they are expressed as differences between two scores, such as occurs in the value-added approach to assessment. Reporting differences in a student's performance (two different instruments or the same instrument at two different times) requires longer assessments than reporting differences in group means.

As in the case of validity, the purpose of the assessment determines which kinds of evidence and procedures are particularly relevant. In this light, let us pass the plane of the reliability issue through our four purposes.

Purposes 1 and 2: Placement and Certification. The reliability of the examination should be high in part because the decisions are important and in part because the subject of the assessment is the individual student. Consistency of decisions is the type of reliability most appropriate for placement and certification examinations. The question is whether the same decision would be made if the student were administered another assessment like the first. Procedures for estimating decision reliability are referenced in end note 25. If this proportion is too low, the examination should be lengthened, particularly by adding items that discriminate at the cutscore. Alternatively, those students scoring near the cutscore could be administered additional questions to improve faculty confidence in its decision.

A lengthy assessment instrument is also suggested in the certification context since the institution has some stake in avoiding false positives. With multiple opportunities to take the assessment, a student who does not have the requisite level of competence is more apt to pass an unreliable assessment at least one of the times it is attempted than to pass a reliable assessment instrument.

Purposes 3 and 4: Course, Program, and Institutional Evaluation. Because data for groups of students (rather than for individual students) are being reported, reliability is of less concern. It is true that if a matrix sampling plan is used, results could be different not only because of measurement error (e.g., using a different sample of items) but also because of sampling error (using a different sample of students). Nevertheless, assuming a representative sample of students responds to the evaluation instruments, scores computed by aggregating responses to several items will be sufficiently precise for most purposes, even with samples as few as 25 or 50. (Such aggregation, however, comes at the price of less detailed information.) Suppose, instead, one wished to report the results for a single item, such as what percentage of students would answer "yes" to question 13 if asked. As indicated previously, a random sample as small as 100 students, regardless of how many are in the course, program, or institution, is sufficient to answer such a question with a margin of error of 10 percent or less.

Value-added claims will be somewhat less reliable because they are based on the difference of two measures, but even so, representative samples of a few hundred are more than sufficient.

What Assessment Results Will Be Reported?

A major value of a college assessment is the information it generates. This information can be conveyed in different forms. How much detail is needed? How should the results be expressed? What comparisons are important?

Concepts and Issues

Diagnostic Information. The more scores that are reported, the more diagnostic the information. Fine-grained reporting can be helpful to the student and the institution. Knowing strengths and weaknesses, highs and lows, can be an aid in studying, planning a curriculum, and communicating the assessment findings. The ability to provide diagnostic information, however, comes at a cost. Information based on only part of an examination can be quite unreliable and misleading; complicated score reports can be misunderstood. Assessments need to be noticeably longer if the detailed results are to be reliable.

Reporting Formats and Metrics. Results can be expressed as number correct (raw scores), units of proficiency, or in some type of norms such as percentile ranks or standard scores. Raw scores and units of proficiency are most useful for well-structured domains in which the institution wishes to focus on a comparison between the performance and some standard. Criterion-referenced interpretations are those in which results have meaning in reference to a standard rather than in reference to how other students did. Examples are: the student can type 45 words per minute, can write an essay at Level 3 (where examples of Level 3 writing are shown), or can carry on a simple conversation in French.

In contrast, norm-referenced interpretations are those in which examination results are compared to those of a reference group, called a norm group. Commercially-published assessment instruments frequently have norm group data. It is important that the institution using these data be sure that the norm group is one to which it wants its students to be compared. Faculty should also be sure that the instrument is administered according to the procedures specified by the publisher. Unfortunately, many norm groups are poorly defined or consist of student volunteers with unknown characteristics, thus preventing intelligent interpretations of the results. As an alternative, the institution can establish its own norm group, for example, a current or previous class.

Comparative Analyses. Any information can be analyzed a great many different ways. Typically, any of several procedures and analyses are valid. In these days of computers, workers no longer have to limit themselves to one or a few analyses. For example, if one is uncertain whether or not to include certain students in a particular analysis, the data can be analyzed both ways.

Comparisons hold much interest. For example, comparisons of this year's data with data for previous years; comparisons among groups of students, programs, and institutions; and relationships between selected individual, departmental, and institutional variables and the assessment results may all be of interest. Adverse impact statistics are frequently sought. The general principle is that the data being compared should be contributed by groups that are as much alike as possible except on the dimension of interest. Thus, for example, if the data from two different years are being compared, the students in both groups should meet the same criteria for being included (e.g., only juniors, from the same department) and the instruments should be administered under the same conditions.

As indicated below, the purpose of the assessment will determine how important diagnostic information and norms are and what comparisons are of most interest. In any report, the scores of individuals identified by name should be revealed only with their informed consent or only to those with a legitimate professional interest in particular cases.[26]

Purpose 1: Placement. Because the purpose of the examination is to place students into appropriate courses or programs, the availability of diagnostic information and norms is of secondary importance. Nevertheless, many examinations of basic skills yield diagnostic information and are recommended for use in placement decisions. Of course, providing all students with a full profile of their strengths and weaknesses would be helpful, and offering some of that information might be possible. The content of a placement examination is not intended to sample all aspects of the subject matter, and those it does include are not weighted equally. The items are chosen for their discriminatory value, not because they are representative of items measuring a particular skill. The position taken here is that a single, fixed-length examination cannot be optimum for both the placement and diagnostic functions.

If the student retains the option of which course or program to select, and the placement examination functions primarily as a guidance tool, then the student should be provided sufficient information to make an informed decision. Such information might include a table showing, separately for groups of students having similar examination scores, the grade distribution for these students in each course under consideration. Needless to say, analyses of the background and subject matter preparation of those students who pass and those who fail the placement examination also would be of particular interest.

Purpose 2: Certification. In addition to reporting the student's score and the cutscore, it would be especially helpful to the failing student if diagnostic information were presented. One form of this diagnostic information that does not compromise the effectiveness of the examination as a certifying instrument is to indicate the number of items associated with each major category of the ability domain and the number of these items the student answered correctly. Norms have little value in this context.

The institution should assemble data on the pass rate for first-time examination takers and for all examination takers, regardless of the number of previous attempts. The second rate will be higher than the first. These pass rates should be shown separately by class year, gender, major racial and ethnic groups, and discipline or major.

Purpose 3: Course and Program Evaluation. A goal of the evaluation is to acquire a reasonably complete picture of the effects of a course or program. Reporting the results in as much detail as the audience is willing and able to comprehend is consistent with this goal. Disaggregated information is especially valuable for curriculum and instructional planning. Differences among subgroups of students, correlations within the data set and with other variables, and comparisons with previous years' data or with available norms can be informative. Comparisons among programs within an institution, unless carefully designed and executed, can be invidious and breed corruption if used selectively to reinforce previously set, politically-determined conclusions.

Purpose 4: Evaluation of the Institution. Detailed reporting should benefit the institution. Although the amount of detail can vary for different audiences, a broad assessment has the potential to provide diagnostic information for the institution. On which outcomes and for which groups were the results encouraging? Disappointing?

When norms are available, they, too, can be informative. Noninstitutional norms may be needed to communicate to interested publics how the institution is doing. A full description of any norm samples employed should be provided. When available, comparisons with outcome data from a previous class can be illuminating. Comparisons with other institutions can be misleading. Student performance depends partially on factors that the institution can do little to change, such as student intellectual aptitude.

A comprehensive evaluation of an institution offers numerous choices for analyses. The assessment data can be supplemented with the growing body of institutional information available in various national data banks. In fact, the institution may wish to store its assessment results in a data bank for secondary analyses by faculty and students.

Summary

This paper has focused principally on examinations and other instruments used in college assessments. It raised a number of critical questions--what to measure, how to measure it, how to know if the measurement is any good, and how to report the results--not for purposes of providing a textbook on measurement, rather to winnow out the concepts and issues that are particularly applicable in higher education settings. Toward that end, I have stressed the practical implications of these issues in each of four major purposes that govern higher education assessment programs: placement, certification, program evaluation and institutional evaluation. This presentation has been intentionally general, and has served to establish a framework for the subsequent essays in this volume that deal with specific areas, problems, and technologies of assessment.

Decisions informed by assessment results can affect the life and spirit of students, courses, and colleges themselves. If those decisions are to be beneficial, the results of assessments must be of a quality worthy of their importance. It is toward that end that this essay has sought to guide the reader through the questioning process.

End Notes

1. The discussion in this chapter is limited to tests (including performance measures), surveys, and student-constructed academic products. Excluded are a myriad of other indicators of the value of a college program, including measures of cost, enrollment data, testimonials about the reputation of the institution, earnings of graduates, and so on.

2. Other compatible purposes include encouraging faculty to examine their curricula, meeting a State mandate, and sending a public message about what the institution values.

3. Although program evaluations or institutional assessments may not purport to evaluate faculty, they may be perceived to do just that. Fear that the assessment is a covert faculty evaluation mechanism can result in a low degree of faculty cooperation.

4. For an excellent introduction to performance assessment, together with practical guidelines, see Richard J. Stiggins, "Design and Development of Performance Assessments," Educational Measurement: Issues and Practice, vol.6, no.3 (1987), pp. 33-42.

5. The MAPS (Multiple Assessment Programs and Services) instrument contains subtests that vary along this continuum. For further information, write The College Board, 45 Columbus Ave., New York City, N.Y. 10023-6917.

6. The Academic Profile contains items measuring one of four general skills (college-level reading, college-level writing, critical thinking, and using mathematical data) within one of three broad subject areas (humanities, social sciences, natural sciences). For further information, see John Centra's essay in this volume.

7. How MAPS Can Help You with Placement. New York: The College Entrance Examination Board, 1980, p. 6.

8. College BASE (Basic Academic Subjects Examination) is a standardized, college assessment instrument that clearly specifies the content domain. The instrument describes content in terms of learning outcomes based on a 1983 statement by the College Board of the academic preparation needed for college. For further information, write the Center for Educational Assessment, University of Missouri, 403 S. Sixth Street, Columbia, Mo. 65211.

9. Recognizing the colleges want measures of the outcomes of instruction in the disciplines that would not be as difficult as the Graduate Record Examination Subject Area Tests, the Educational Testing Service and the Graduate Record Examinations Board have constructed a set of new examinations based on the GREs. The sponsors claim that these examinations are less difficulty, appropriate for all seniors majoring in a field, more convenient to administer than the GRE Area Tests, and hence more appropriate for program or institutional evaluation. For further information, write Major Field Achievement Tests, 23-P, Educational Testing Service, Princeton, N.J. 08541-0001.

10. Other sources include expert judgments about the merits of course materials (syllabus, assignments, class handouts, examinations, textbooks, etc.) enrollment and attrition figures, costs data, and facilities.

11. Derek C. Bok, Higher Learning. Cambridge: Harvard Univ. Press, 1986, p.59.

12. Write to the Assessment Resource Center, University of Tennessee, 2046 Terrace Ave., Knoxville, Tenn. 37996-3504.

13. The Institution published the Mental Measurements Yearbooks, which include comprehensive descriptive information and critical reviews of commercially published tests. The Institute also offers an online computer databse service through BRS Information Technologies, that provides monthly updates in between publication of the Yearbooks. The label for the database is MMYD. Further information and announcements may be obtained by writing the Buros Institute, University of Nebraska, 135 Bancroft Hall, Lincoln, Neb. 68688-0348.

14. Write to the Test Corporation of America, 330 W.47th Street, Suite 205, Kansas City, Mo. 64112.

15. See Emily Fabiano and Nancy O'Brien, Testing Information Sources for Educators. ERIC TME Report 94. Princeton, N.J.: Educational Testing Service, 1987. Includes a listing of printed material and computer-based sources of test bibliographies, agencies providing test information, locations of major and regional test collections, and names and addresses of test publishers. Write to the ERIC Center, American Institutes for

Research, 3333 K Street, N.W., Washington, D.C. 20007.

16. Richard Light, personal communication, Dec. 7, 1987.

17. American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, D.C.: American Psychological Assoc., 1985. See especially Sections 13 and 14.

18. A good introduction is G.L. Marco, "Equating Tests in an Era of Test Disclosure." In B.F. Green (ed.), New Directions for Testing and Measurement, No. 11. San Francisco: Jossey-Bass, 1984, pp. 105-122. Further introduction and details can be found in W.B. Angoff, Scales, Norms, and Equivalent Scores. Princeton, N.J.: Educational Testing Services, 1984. The 1984 publication is a reprint of Angoff's classical, but very relevant treatment of the topic in 1971.

19. An elementary introduction can be found in S.A. Livingston and M.J. Zieky, Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. Princeton, N. J.: Educational Testing Service, 1982. Other key references on this topic include L.A. Shepard, "Standard Setting Issues and Methods," Applied Psychological Measurement, vol. 4 (1980), pp. 447-467, and R.A. Berk, "A Consumer's Guide to Setting Performance Standards on Criterion-Referenced Tests," Review of Educational Research, vol. 56 (1986), pp. 137-172. The latter reference identifies and rates 38 methods.

20. See R.A. Berk, (ed.), Handbook of Methods for Detecting Test Bias. Baltimore: Johns Hopkins Univ. Press, 1982.

21. A brief introduction to current thinking about test validity can be found in Section 1 (pp. 9-18) in the Standards (see end note 17), as well as in Howard Wainer and Henry Braun (eds.), Test Validity. Hillsdale, N.J.: Lawrence Erlbaum Assoc., 1988. A thorough treatment of the topic will appear in R.L. Linn's edited volume, Educational Measurement (3rd edition), which will be published by Macmillan in 1988.

22. A subscale of the College Outcome Measures Project (COMP) of the American College Testing Program. For more information, write to ACT COMP, P.O. Box 168, Iowa City, Iowa 52243.

23. See George F. Madaus, "Minimum Competency Testing for Certification: The Evolution and Evaluation of Test Validity." In G.F. Madaus (ed.), The Courts, Validity, and Minimum Competency Testing. Boston: Kluwer-Nijhoff Publishing, 1983, pp. 21-61.

24. See Jason Millman, "A Checklist Procedure." In N.L. Smith (ed.), New Techniques for Evaluation. Beverly Hills, Calif.: Sage Publications, 1981, Vol. 2, pp. 309-314 and 316-318.

25. A brief introduction to current thinking about test reliability can be found in Section 2 (pp. 19-23) of the Standards (see end note 17). A thorough treatment of the topic will appear in Linn's forthcoming 3rd edition of Educational Measurement (see end note 21). Traditional formulas for reliability and the standard error of measurement can be found in any elementary textbook on educational measurement. One notable review of methods for determining decision consistency is that of Roos E. Traub and Glenn L. Rowley, "Reliability of Test Scores and Decisions," Applied Psychological Measurement, vol. 4 (1980), pp. 517-545. A computationally simple procedure for estimating decision consistency with only one test administration is offered by Michael Subkoviak, "A Practitioner's Guide to Computation and Interpretation of Reliability Indices for Mastery Test," Journal of Educational Measurement, in press (1988).

26. See Section 16 (pp. 85-87) of the Standards (see end note 17).

Diverse and Subtle Arts
Assessing the Generic Academic Outcomes of Higher Education

by Leonard L. Baird

This chapter is concerned with the assessment of generic academic outcomes of higher education. The first question we need to address is "What are these outcomes?" That is, what are the generic cognitive operations that one would expect or hope to see developed or improved through a college education, particularly through its general education components? When asked such questions, most professors and administrators respond with the ideas that such outcomes or cognitive components should not be tied to knowledge of a particular subject matter but that students should be taught to think for themselves, reason deductively and inductively, demonstrate critical thought and solve problems in general situations. I would surmise that most educators feel that they can define these terms. However, when pressed to do so for general education or the college curriculum, they encounter unexpected difficulty. As Cuban (1984; p. 676) points out, defining these operations ". . . is troublesome to both social scientists and practitioners. Troublesome is a polite word; the area is a conceptual swamp."

After spending years on the issue of definition Cuban decided that

> ". . . reasoning, thinking, critical thought and problem solving as notions in the minds of teachers, administrators and parents were for the most part indistinguishable. In the general discourse that I participated in, these phrases were interchangeable. I also found that the most frustrating, enervating discussions I engaged in tried to define reasoning, often breaking down into semantic hassles that would warm a Talmudist's heart." (p. 677).

Whatever the definitional problems, a good deal of the recent discussion about assessment has been concerned with these generic academic outcomes.

Although there is a large conceptual muddle concerning the outcomes of higher education, it is possible, and ultimately useful at this point, to distinguish among three interrelated types of variables: basic skills, "general learned abilities," and generic academic outcomes. Although these distinctions are based as much on measurement practice as on real distinctiveness, they can be viewed separately. "Basic skills" are usually considered to be those fundamental skills which are prerequisite to typical college work, such as the rudiments of arithmetic and (sometimes) geometry, and the ability to read an ordinary paragraph and understand it. The emphasis may be placed on skills in the sense that they are the academic tools needed to begin academic work. Tests of "basic skills" are often used to determine whether a student is ready for college work or requires additional work in courses designed to enhance these skills.

"General learned abilities" is a phrase often used in the publications of Educational Testing Service (ETS) to describe the variables their tests measure, usually termed verbal and mathematical or quantitative ability. These are supposed to be broader and more general than basic skills and are supposed to represent the abilities that underlie most academic work. That is, unlike basic skills, which are needed to begin study, "general learned abilities" are supposed to be related to performance across the curricula. The higher a student stands on these abilities, the more difficult the work that can be attempted, and the more successful the student in that work. At various times, ETS has described these abilities in terms that suggest that they are almost permanent and unchangeable. In any case, they are usually measured at two educational transition points: application to college, by the Scholastic Aptitude Test (SAT), and application to graduate or professional school by the Graduate Record Examinations' (GRE) General Test, the Graduate Management Admissions Test (GMAT), etc. They are thus designed to assess the student's readiness to take on further work. Given their purpose, these measured abilities are deliberately not tied to any particular pattern of coursework. Thus, the GRE General Quantitative problems require no formal mathematics courses beyond high school algebra and visual geometry. (The GRE will be discussed further below.) These "learned abilities" may be so stable as to be unaffected by educational programs, and thus would be unsuited as criteria of learning.

Generic academic outcomes, although founded on basic skills, and dependent to some degree on broad "learned abilities," are those outcomes which are assumed or hoped to develop as a result of the academic work pursued in college. These generic outcomes are thought to come about, not through any particular course, but as the product of a variety of educational experiences which all contribute to the outcome. These would include such outcomes as the ability to draw inferences, to use various modes of reasoning properly, to make logical arguments, and to analyze situations and problems. Although there are great difficulties in identifying these outcomes (just ask anyone who has served on a committee on general education requirements), they can be seen as reflections of the broad educational goals of our curricula and extracurricula. They most often center on critical thinking, reasoning, and problem solving.

I would argue that these generic academic outcomes reflect the core of college teaching and curriculum much more closely than either basic skills or general learned abilities. Similar arguments have been made by Woditsch (1977), Warren (1978) and Ewens (1979). The rest of this essay will be concerned with the procedures needed to assess these outcomes.

As this volume concerns assessment, it should be noted that the number of available instruments that are designed to assess general critical thinking, reasoning, and problem solving is quite limited. For example, "critical thinking" has most often been assessed by one instrument, the Watson Glaser Critical Thinking Appraisal. "Reasoning," interestingly enough, has seldom been assessed directly as a cognitive outcome of academic education. Instead, it has most often been approached through measures based on the theories of Kohlberg and Perry concerning moral, ethical and intellectual development, according to which students may move from dualistic "black-white" thinking, through relativism, to commitments based on principles.[1] Because of their purpose, these measures are only indirectly related to what most academics think of as reasoning. There are virtually no measures designed to assess general problem solving other than a few experimental game-like exercises used by researchers attempting to use computers to study thinking.

Because the greatest amount of experience and research in these three areas has focused on assessments of critical thinking, this paper will use that as an example in outlining what the author believes is the most reasonable approach to assessment of generic academic outcomes, which is to think of the task as analogous to professors' classroom assessments of student learning.

How Assessment and Good Teaching Practices are Alike

Good teaching involves establishing clear objectives, defining criteria to judge performance, developing methods to assess that performance, and using the method. The results provide feedback to learners and teachers so that learners can judge their own mastery of the objectives, and teachers may judge their success in teaching. These are ordinary and expected aspects of college classwork.

The current move to assess college outcomes can be seen as an extension of these typical academic procedures to the overall undergraduate experience. They can help us assess the qualities that are important in academe, such as critical thinking, imagination, and sensitivity. Although the development of these characteristics is often stated as a goal of higher education, the insistence of both faculty and legislators on better means of assessment indicate that we fall short of the goal in the eyes of many. However, the defense of colleges and their programs is not the most important use of general assessments of outcomes in higher education, rather it is that comprehensive assessments can lead to a healthy discussion of overarching academic goals, curricula, requirements, and, perhaps most important, the process as well as the content of learning. Thus, it would seem that the improvement of learning in a college would be dependent on institution-wide assessment, particularly when broad academic outcomes are considered. That is, as valuable as it is to assess students' mastery of specific courses and particular majors, it is very difficult to "add up" the evidence from these bits of information to know how much a student has gained from college, or to assess how well the college is educating the "whole person."

If we accept this argument, we need to assess the generic academic outcomes of higher education in order to understand the quality of undergraduate learning. How may we do so in an educationally and technically responsible way? I would like to argue that we can view this task as an extension of the procedures good professors use in assessing the learning of students in their courses. That is, responsible psychometricians follow procedures that are very similar to those used by responsible professors.

1. The Establishment of Objectives and Setting Criteria for Judging Performance. In practice, psychometricians follow procedures that are formalized analogues to setting course objectives. Just as professors should decide and specify what they want their students to gain from their courses, psychometricians attempt to "define," rationally and systematically, the content, type of assessment, and level of performance to be included in their instruments. These "test specifications" are usually developed with the advice of content specialists who, in essence, supply a model of knowledge, skills, and competencies in the discipline under concern. For example, when the Graduate Record Examinations Board develops a new version of its test in chemistry, it includes a panel of chemistry faculty drawn from different institutions and specialties so that a diversity of programs is represented. The panel then defines the substantive content to be covered, such as properties and reactions of the important elements and their compounds, biochemistry, acids and bases, chemical equilibrium, etc. The weight to be given to facts, principles and techniques is determined, thus providing a rough model of the structure of the discipline. The GRE example indicates that this kind of exercise can be conducted within disciplines, but can it be done across disciplines to include generic academic outcomes of undergraduate learning?

For a college to follow these procedures in developing objectives or "specifications" for generic academic outcomes would require it to set up a committee including faculty from a diversity of disciplines, who teach both introductory and advanced courses, and who teach students who are majors and non-majors. These faculty would need to reach some consensus about the content, principles and intellectual skills they believe are fostered by the institution's programs. Since there are limits on the time that can be devoted to assessment, some decision about the weight to be given to different content areas would need to be made. Plainly, the difficulty of reaching consensus on these questions would vary by type of institution; most small traditional liberal arts colleges would presumably have fewer difficulties than large, multipurpose institutions.

As long as we stick with general categories of outcomes it is quite possible to gain general agreement about goals. As Bowen (1977) notes, "if the goals of education are defined as a list of desirable objectives, without priorities among them, there is even considerable agreement" (p. 32). Although agreement on general types of outcomes may be reached, the agreement may quickly break down when specifics are considered. The general goal of "critical thinking" provides a useful example. The most common academic conception of critical thinking is that of the ability to criticize claims or propositions in terms of reason rather than emotion, or the authority or tradition that makes them. In this conception, critical thinking is primarily an attitude, a stance of persons who are willing to question their own and others' ideas, and who, relying on their own logic and understanding, are willing to change their views.

However, some have argued strongly that this attitude of autonomy and skepticism is insufficient for critical thinking; the thinker also needs intellectual tools which allow the analysis and evaluation of intellectual claims. In Teaching Values in College, Morrill has recommended that students learn to analyze the validity of moral claims according to the criteria of their consistency, coherence, comprehensiveness, adequacy, duration, and openness. More traditionally, courses in logic emphasize such tools as identifying the steps in an argument, detecting assumptions and ambiguities, considering alternatives and avoiding common errors in argument (Ennis, 1986). The assumption is that these techniques can be taught much as writing can be taught. It is unclear, however, whether people can apply such techniques uniformly. For example, students may be able to synthesize and evaluate arguments in their major field, but be unable to show more than mere comprehension in another area.

More fundamentally, it is unclear as to whether skills in general logical reasoning and "critical thinking" are the same.[2] Recent philosophers such as Toulmin (1976) and McPeck (1984) have argued that the criteria for a valid argument are different in different fields, and that training in logic or general intellectual heuristics will therefore not necessarily result in critical thinking in specific disciplines. The problem, then, is not that students cannot transfer their thinking from one discipline to another but that what constitutes "thinking" may differ from one area to another. Thus, in order to "think well" in a discipline, a student needs to understand the structure of knowledge in that area; to think well across disciplines, a student needs to learn the structure of knowledge in a variety of disciplines. But as Geertz (1983) has suggested, it may be impossible or even undesirable to expect that different disciplinary and practical ways of thinking will resolve into a single kind of "critical thinking." Furthermore, even if we could develop a general scheme of thinking, the way this thinking is used in any particular situation is largely dependent on the context. As Geertz points out, "everything is general in general but particular in particular." These philosophical points are reinforced by studies of thinking and problem solving (Chipman, Segal, and Glaser, 1985), which suggest that different kinds of problems or situations require strategies that are specific to those situations. The most generally applicable procedures are also the weakest and least efficient.

In sum, there are practical and theoretical reasons why it may be very difficult to obtain consensus on educational objectives in these subtle areas, except in the most general way. As illustrated by "critical thinking," when more specific objectives are defined, any earlier consensus may dissolve. Furthermore, (a) such generalized thinking may not exist or (b) if it does exist, it is so weakly manifest in general situations as to be nearly useless. This problem of generality applies to other general goals identified by Bowen (1977): intellectual tolerance, aesthetic sensitivity, and creativity.

2. Developing Methods to Assess Performance. The difficulties in defining goals and criteria to judge performance have many implications for methods of assessing performance. What sort of method or instrument would assess the multiple approaches to cognition and reasoning found in different disciplines? To use a concrete example, what would we consider evidence that a student had reached an adequate level of critical thinking? What sort of exercise would represent a mode of critical thinking which could be generalized across the curriculum? Again, these sorts of questions are very similar to those the psychometrician faces when developing an instrument. That is, given the content and skill specifications for the examination, the psychometrician must devise specific questions or exercises that assess the knowledge and skill in question. For example, a potential exercise in chemistry might describe an experiment and ask students to indicate what it is about (knowledge), and how the procedures were used (skills). In the case of such general outcomes as critical thinking, problem solving, and the like, it is very difficult to develop instruments which are applicable across the curriculum (McMillan, 1987).

In addition, just as in the classroom test, the methods used determine the kinds of variables that can be assessed. For example, Table 1 (Taken from Gronlund, 1986) shows some of the

Table 1. Types of Complex Learning Outcomes Measured by Essay Questions and Objective Interpretive Exercises

Type of Test Item	Examples of Complex Learning Outcomes That Can Be Measured
Objective Interpretive Exercises	Ability to— identify cause–effect relationship identify the application of principles identify the relevance of arguments identify tenable hypotheses identify valid conclusions identify unstated assumptions identify the limitations of data identify the adequacy of procedures (and similar outcomes based on the pupil's ability to *select* the answer)
Restricted Response Essay Questions	Ability to— explain cause–effect relationships describe applications of principles present relevant arguments formulate tenable hypotheses formulate valid conclusions state necessary assumptions describe the limitations of data explain methods and procedures (and similar outcomes based on the pupil's ability to *supply* the answer)
Extended Response Essay Questions	Ability to— produce, organize, and express ideas integrate learnings in different areas create original forms (e.g., designing an experiment) evaluate the worth of ideas

Source: Gronlund (1985)

types of complex learning outcomes that can be assessed by different types of tests. "Objective interpretive exercises" consist of series of objective items based on a common set of data or stimulus materials: statements, paragraphs, tables, charts, graphs, or pictures. Items which ask questions about the material can capture students' ability to recognize the relevance of information, recognize warranted and unwarranted generalizations, apply principles, recognize assumptions, recognize inferences, and interpret graphs, among other skills. However, since it is based on selecting the correct answer, the interpretive exercise is confined to learning outcomes at the recognition level. It also is confined to specific aspects of cognitive processes, so it does not assess students' ability to integrate these processes when faced with a particular problem.

The "restricted response essay question" limits the content and response. An example would be "State the main differences between the Korean War and the previous wars in which the United States has participated. Answer in a brief paragraph." Gronlund considers the restricted response question to be most useful in measuring learning outcomes requiring the interpretation and application of data in a specific area, such as the ability to **formulate**, rather than recognize, a valid conclusion. Students **supply** an answer. The "extended response essay question" asks a question that allows for much greater freedom of response to select, organize, integrate, and evaluate ideas. An example would be "Evaluate the influence of Freud's theories on the development of psychology as a science." Although these kinds of essay questions can assess complex cognitive processes, they are difficult to score, and the processes to be assessed need to be clearly defined.

The approach taken by the authors of most of the measures attempting to assess critical thinking is to use objective interpretive exercises. The most commonly used instrument, the Watson-Glaser Critical Thinking Appraisal uses multiple choice questions to assess induction or inference. This performance requires the student to distinguish the probable degree of truth or falsity of inferences drawn from a statement or situation, recognize unstated or presupposed assumptions in a statement, deduce a particular conclusion from two syllogistic statements, interpret whether various conclusions follow from the description of a situation, and evaluate the strength of arguments supporting some proposition.

Although these cognitive operations are important, they assess only some of the skills often identified as characterizing critical thinking. For example, Ennis (1986) also includes the critical thinking "skills" of focusing on a question, seeing the structure of an argument, asking questions of clarification and/or challenge, judging the credibility of a source, judging the relevance and probable accuracy of reports and criteria, judging the generality of data or information, seeking evidence and proposing hypotheses, making value judgments, defining terms, and planning and deciding on an action to be taken. (Some of these are assessed in the Cornell Test of Critical Thinking.) In addition, Ennis lists 14 "dispositions" or dimensions of the critical thinking attitude that are needed for students to use their critical thinking abilities or skills. However one may stand on these issues, the point is that the method one uses limits the range of skills one can assess, and that the chief method of most currently available assessments of critical thinking, the multiple choice test, limits their range to a set of discrete technical skills.

For example, although the publications of the GRE Board specifically warn against using the Graduate Record Examination General Test for any purposes other than admission to graduate study and guidance, it has been proposed, and in a few cases, used, as a measure of the general outcomes of college. Is the GRE Board right to be wary of such usages?

First, let us remember that the GRE is basically an admissions instrument, oriented to identifying graduate school applicants who are likely to succeed in graduate work in a wide variety of disciplines, a purpose quite different than assessing the educational progress of students. Secondly, it is designed to assess very broad abilities developed over a long period of time. In fact, the correlation of GRE scores with earlier SAT scores are nearly as high as one would obtain by readministering the SAT. As a result, it is very difficult to demonstrate any relative gain due to teaching or curricula. Third, the range of variables assessed by the GRE is actually fairly narrow. It does not assess such well known "developed abilities" as spatial, mechanical, and clerical abilities, which are also involved in many curricula.

Even within the realm of "verbal" and "quantitative" abilities, the definition is limited to the types of items used. For example, the Verbal section consists of items based on reading paragraphs similar to those found in textbooks, choosing words that best complete sentences, solving analogies, and identifying antonyms. It would appear that these items assess vocabulary and reading comprehension, but there is, in fact, no explanation of what verbal abilities they are intended to assess. Likewise, although one could obtain sub-scores on the quantitative items that focus on interpreting data, or comparing quantities, or basic math, the substance of these scores is unclear. For example, quantitative comparisons include items based on knowledge of routine mathematical symbols, basic geometry, understanding of word problems, basic arithmetical and algebraic procedures, and practical math, all mixed together.

The Analytic section of the GRE consists of reasoning items which test examinees' abilities to recognize correct deductions and inferences about structural relations among a set of objects, and logical reasoning items which test examinees' abilities to identify correct and incorrect deductions from stated arguments. Although potentially useful, many of the analytical reasoning items seem more like complicated word puzzles than assessments of the ability to analyze, and the logical reasoning items focus on solving single logical problems, rather than assessing the overall logic of an argument. Furthermore, there is no evidence as to whether GRE Analytical scores rise or fall with instruction that should presumably affect them. Perhaps most importantly, the Analytical score does not require the production of logical, reasoned arguments.

It is easy to understand why the developers of commercial tests would use the objective multiple choice method. Other methods are more difficult to score, more costly and more time consuming both to administer and score. The same pressures toward multiple choice examinations would apply to assessments of other general cognitive outcomes of higher education. Thus, no matter how clearly the items are written, these assessments would be limited to a subset of the aspects of the outcome, which can be recognized and selected by the student. To assess students' cognitive ability to define problems, to formulate hypotheses, to obtain or organize data, and to draw conclusions we must turn to the costlier and more difficult procedure of essay examinations.

3. Using the Assessment Method. Professors give their examinations to their classes and judge from students' responses whether the questions accurately obtained the information sought, i.e., whether their questions were good ones for assessing the learning of students. The procedures of psychometricians are more elaborate extensions of this process. In developing a test, the psychometrician looks for reliability or consistency of response, validity or the extent to which the test is related to criteria of interest, and, for educational tests, the tests' sensitivity to change. It is important to note the relationship between method and reliability. The complexity and thereby the variability in scoring increases as the degree of free responses increases. Thus, the "objective" multiple choice test is usually more reliable than the limited free response method, which is more reliable than the essay question method, even though one can increase the reliability of essay exams considerably by carefully defining objectives, properly framing the questions, having clear scoring values, and by providing training in scoring (Gronlund, 1985). There are thus additional reasons for psychometricians to favor the objective multiple choice examination even though it may not be the best method for assessing many types of educational objectives.

Professors share with psychometricians a concern for the validity of their examinations, that is, whether the exams accurately assess what they wish to know. Again, the psychometrician uses more elaborate and standardized versions of what professors do. After the test, comments from students may tell the professor how well his or her test assessed the material to be learned, and the consistency of the results with other information about the student's performance may suggest how well it assesses learning. Analogically, the psychometrician may ask panels of experts to judge the extent to which the test tasks meet the test specifications in order to see how well they represent the domain of tasks to be measured ("content-validity"). A second procedure is to compare performance on the test with performance on other tests that are purported to measure the same area ("concurrent validity") and a third is to see how well test performance predicts future performance ("predictive validity").

If we correlate one test of critical thinking with another, all we know is that they give consistent results. And we may begin to wonder about the conceptual justification of the test if it correlates rather highly with paper and pencil tests of intelligence, as the Watson-Glaser does. To what other kinds of instruments would we expect a test of critical thinking to be related? An even more difficult question is "What future performance would we expect a test of critical thinking to predict?" That is, what behavior or accomplishment would suggest that a person was a "critical thinker"? Currently available tests of critical thinking have chiefly related scores to convenient criteria such as college grades, and there have been few attempts to define more clearly relevant criteria. This strategy may be the result of the difficulty of even thinking of some real-life, socially relevant criteria for critical thinking. For most of the areas of generic academic outcomes, it is difficult to devise "real-life" criteria beyond the claims of the test itself.

Psychometricians also judge the validity of a test in terms of the meaning of the scores as a measure of the outcome it purports to measure. One method is to ask test takers to "think aloud" as they respond to the test materials, in order to analyze the mental processes required. This method can provide information about what the tests really measure. For example, when this has been done with some tests that are supposed to measure knowledge or "reasoning," some items seem to reflect the intended knowledge or reasoning process while others can be solved by features of the items (Greeno, 1978; Gronlund, 1985).

Another method is to compare the scores of known groups, e.g., for tests of critical thinking, entering freshmen, seniors, graduate students, recipients of PhDs and professors. The assumption is that scores should rise with increased education thus demonstrating the "validity" of the test. Another strategy would be to compare the scores of groups that presumably use the skill in question. In the case of critical thinking, we would expect the scores of judges and patent attorneys to be higher than the scores of the general public. These comparisons have rarely been pursued principally because of difficulties in obtaining the cooperation of respondents.

One can also correlate the test with a wide variety of other tests. Naturally the psychometrician looks for higher correlations with similar tests and low correlations with dissimilar tests. Here the effort is to show that the test assesses a common underlying variable. If the test does not correlate with similar tests, it does not appear to be measuring the same thing. On the other hand, if it correlates highly with a supposedly different kind of test, one may doubt whether it is measuring the variable it is purported to, or whether it is really a measure of something else. One of the criticisms of the Watson-Glaser, for example, is that it correlates rather highly with tests of reading and vocabulary. Thus it may be measuring the ability to read carefully as much as it measures something called "critical thinking."

Finally, professors use their tests to show that their students have learned the material in their classes. Obviously the tests need to reflect the content of the classes. Students at the beginning of the course would not be able to pass the final examination. More generally, students who have been taught the material would be expected to know it; students who have not would not be expected to know it. Almost all students at the beginning of a course in Russian would not know the Cyrillic alphabet; students at the end should be able to read some Russian prose and write a paragraph in Russian. Test scores would reflect these differences dramatically. Although the differences in scores in courses where students have some prior knowledge (such as precalculus or United States economic history) would be less dramatic, they would show that students have learned while enrolled in the course. In short, the tests are sensitive to change, and thus will assess students' improvement in learning. In addition the test results can lead to actions, i.e., if a student does not perform well on a part of a midterm, the student can review the material. If all the students do poorly, the professor may spend a class period presenting the material again, probably in a different way.

In the case of general measures of such qualities as critical thinking or problem solving, it is much more difficult to show improvement, or to link improvements in scores to any particular curriculum, course, or teaching strategy (McMillan, 1987). That is, scores on such broad measures will be influenced by many factors, not the least of which is the students' level of critical thinking at the beginning of a program. Thus, although seniors score higher than freshmen on the Watson-Glaser, the great majority of studies of particular programs designed to enhance critical thinking showed no significant incremental gain (McMillan, 1987).

More basically, as Pascarella (1985) notes, there are many possible reasons for an increase in a score other than the college experience. One is maturation, whereby individuals' scores increase with age, whether or not they attend college. Another is that students who do not gain tend to drop out, so that the only seniors left to test are those who have gained. However, the real reason for the apparent lack of influence by various academic attempts to affect positively such qualities as critical thinking may lie with the generality of the instruments used. That is, when instruments such as the Watson-Glaser correlate highly with general intelligence and general reading skill, it may be quite difficult to show change. More specific measures of critical thinking which are related to disciplines, such as those used by Dressel and Mayhew (1954), show more changes specifically related to course work.

As part of their evaluation of general education in 19 colleges, Dressel and Mayhew, working with faculty committees, developed a Test of Science Reasoning and Understanding, a Test of Critical Thinking in the Social Sciences, and in the humanities a procedure, Guide to Critical Analysis in the Humanities. A set of skills in each of these areas was identified. For example in the science reasoning test, there were five abilities: The ability to recognize and state problems; the ability to select, analyze, and evaluate information in relation to a problem; the ability to recognize, state, and test hypotheses, and other tentative explanations; the ability to formulate, recognize and evaluate conclusions; and the ability to recognize and formulate attitudes and take action after critical consideration. There were more specific subskills within each of these skills. For example, the second, dealing with information, requires students:

a) to recognize when the information they possess is inadequate for a given problem.
b) to indicate kinds of sources of information appropriate for a given problem.
c) to evaluate the authenticity of given sources of information in relation to a given problem.
d) to indicate their ability to apply information they possess or have gathered to the solution of a given problem.

The specific items were based on content from physics, biology, chemistry, geology, meteorology and ecology. The tests were correlated with other tests, with essay versions of the tests, and subject to various statistical analyses of their reliability and internal structure. Perhaps most importantly, analyses showed that the test scores increased with years in college, and increased after students completed specific courses. One of the most interesting findings was that students who took a combination of specialized and general education courses made larger gains than those taking either type exclusively. Unfortunately, these tests have been out of print for years.

A related problem of sensitivity to change is the "ceiling effect." Some freshmen score high on general academic measures. Thus, when they are retested as sophomores or seniors, they can show little gain. In contrast, the freshmen who score low have plenty of room to increase their scores. It would appear, then, that the more able student had gained little or nothing, and the less able student a great deal. This result has been reported several times in studies of general critical thinking (McMillan, 1987). The result would hold even if the more able student had gained just as much or even more in critical thinking because the tests do not provide a means to demonstrate that gain. Thus, selective colleges enrolling students who already possess the general academic skills measured on the test may have a difficult time showing that they have an impact. Measures directly tied to academic disciplines would be much less subject to this problem.

In sum, one needs to demonstrate that assessments of generic academic outcomes are reasonably reliable, validly related to educational outcomes, and sensitive to the impact of educational programs on students. Although commercially available measures of generic outcomes usually show adequate reliability, they do not often demonstrate their relation to educational outcomes, and are often insensitive to educational change.[3]

4. Providing Feedback to Learners and Teachers. As noted earlier, professors can use their examinations as an integral part of instruction. By carefully pointing out to the student the material that was not mastered or misunderstood, the professor continues to teach and the student to learn. Likewise, when the professor uses the examination to demonstrate to the student that he or she has learned the material, the student can concentrate on learning new material. When most of a class misses a question it is a sign to the professor that further review or presentation of the material in a new way is needed. If most of the class answers a question correctly, the professor can move on to new material with confidence. It is not as simple to engage in this process with measures of general academic outcomes. Since performance on measures of these outcomes can be influenced by many different factors, including the students' living group, peers, extracurricular activities, employment and life experiences, one should probably temper the conclusion that the college has played the major role in determining the score.

When an assessment device uses content that is not specifically related to the goals and programs of the college, the possibility of this sort of misattribution of effects is increased. More importantly, it is difficult to use general measures of abstract qualities to meet the primary purposes of assessment, which according to Warren (1987) are to give students, faculty and administrators information on what has been learned, how well it has been learned, and through what means it has been learned. The measures of generic academic outcomes are of necessity so general and abstract that they do not serve any of these purposes well.

Whither Assessment?

So, where does this leave the educator or administrator who believes in generic academic outcomes, and who believes that colleges should be able to demonstrate that the minds of their students have grown in these terms? I would argue that, given our current knowledge, the assessment of critical thinking, problem solving, and such related outcomes as creativity can best be done within the context of discipline or program.[4]

My argument is based largely on recent philosophy and research on thinking which suggests that these qualities may not exist "in general," free from any context or background. The research, reviewed by Glaser (1984), Perkins (1985), and Chipman, Segal, and Glaser (1985), indicates that knowledge of a particular area is a significant factor in the development of thinking, that many intellectual skills are context specific, and that what are needed are measures specific to disciplinary fields.

On the other hand, authors such as Nickerson (1986), Perkins (1986), Quellmalz (1986), and Sternberg (1986) contend that generalized thinking skills exist, and can be developed by appropriate educational programs. For higher education, Nickerson's conception of critical thinking as a prerequisite for good citizenship, and Paul's conception of "dialogical thinking" as critical thought essential to the acquisition of rational knowledge and passions, may be particularly appealing. However attractive these ideas may be, the problem is that there are currently no well developed methods for assessing them, and their educational implications in higher education are still a matter of controversy. And, as noted in this essay, currently available assessments of general critical thinking, problem solving, etc. are so general that it is very difficult to attach performance on the measures to particular curricular experiences.

But if we look for evidence about critical thinking and problem solving within discipline or program areas, the results will be much more acceptable and meaningful to faculty. They will have much clearer and more specific educational implications, and thus should lead to appropriate changes of emphasis in courses.

If we accept this argument, the question of how these generic academic operations can be assessed remains. One approach to the assessment of students' thinking, problem solving skills and creativity would be to a program-related senior project that would necessitate the exercise of those qualities. The approach differs from the traditional senior thesis in some institutions in that it would emphasize the generic academic outcomes more than disciplinary content, and would require that students be apprised of this emphasis. Either program faculty or external experts in the field could evaluate the project. Of course, these faculty or experts would need to be informed about the methods for systematically judging and evaluating products. There are a variety of techniques, described by Cronbach (1984), Fitzpatrick and Morrison (1971) and Priestley (1982), but at a minimum, the evaluation should be based on a set of definitions of expected outcomes; these outcomes should be directly observable; clear definitions of desired behavior should be provided (e.g. "consistently selected proper equipment in laboratory"); evaluators should be quite familiar with the tasks; and evaluations from several judges should be combined.

A more traditional solution would be to develop departmental comprehensive examinations that would attempt to assess the qualities sought. Some very helpful suggestions for constructing and using comprehensive examinations to improve program quality have been made by Banta and Schneider (1988). In addition to having highly relevant assessment materials, the departments these authors worked with reported that the process of developing the examination forced faculty to focus on common learning objectives for students, encouraged consistency in the teaching of basic courses, led to an emphasis on core competencies throughout the more advanced courses, and encouraged departments to produce a clearer and more logical progression of courses from lower to upper division levels. Perhaps most importantly, although the departments had just begun the process when Banta and Schneider wrote their report, the departments felt satisfied that their examinations captured such qualities as creative thinking, problem solving and the application of principles to real life situations in their disciplines.

Another possibility would be to request the testing agency in the field, whether it be the Graduate Record Examinations Board or a professional group, to release test results by item, or at least by item type. An OERI-sponsored research project at Iowa State is using this procedure.[5] With proper statistical controls, student performance on those items that faculty agree require problem solving, critical thinking or creativity in the major field could be examined as potential indicators of the program's success in engendering those qualities.

Some Prospects for Future Measures

Clearly, there are no completely adequate measures for many cognitive skills. However, several trends in recent research suggest that better measures may be developed in the near future. This prospect is based on the great increase in interest and research on cognitive processes in the past 10 years. By examining the processes by which people actually perceive, understand, manipulate ideas, solve problems, and create, it should be possible to develop measures that can be used to assess and improve students' performance. Although it is probably premature and presumptuous to suggest the specific directions this research might take, several conceptions appear especially promising. One is in the identification of the kinds of cognitive abilities by which individuals manage their own thinking (Segal, Chipman and Glaser, 1985), or as Sternberg (1986) calls them, the executive processes in thinking, which are used to plan, monitor and evaluate one's strategy for solving problems. These include defining the nature of the problem, selecting the components or steps needed for a solution, selecting a strategy for ordering the components, selecting a mental representation of the situation, allocating mental resources, and monitoring the solution. Sternberg believes that each of these can be improved. He is also reported to be developing instruments to assess the components in his model.

Another hopeful development is the study of everyday reasoning in naturalistic situations. This has led such researchers as Ennis (1986) to develop models of the critical thinking process that go beyond formal logic. Ennis has attempted to assess the critical thinking variables in the model through essays (Ennis and Weir, 1985). Other researchers have concentrated on everyday decision-making processes. By focusing on how people actually think, decide, and act, these lines of research should lead to more realistic and relevant assessments of important thinking skills.

Finally, although there are only a few examples available in the published literature, several disciplines or fields are currently conducting research into thinking skills within their own areas. Some examples include mathematics, physics, and medicine (See Tuma, Rief, and Glaser, 1980; Segal, Chipman and Glaser, 1985). One encouraging note is that these different researchers seem to be quite aware of each other's work and criticize and build upon their respective contributions. It seems reasonable to expect that their efforts will increase our understanding and eventually our assessment and improvement of thinking skills.

End Notes

1. General "reasoning" is, as Cuban (1984) notes, extraordinarily difficult to define, and consequently as difficult to assess. Although it is common for examinations in mathematics to include items that supposedly assess "mathematical reasoning," general reasoning, deductive reasoning, inductive reasoning and the like have seldom been assessed. Although there is a considerable body of research that bears on reasoning, especially in children, the only measures that have been used to any extent that bear on general reasoning among college students come from studies of student development. Several have been based on Kohlberg's model of moral development (Kohlberg, 1976), but as Nucci and Pascarella's (1987) review reports, the most important research finding using these measures is that "principled moral reasoning is positively associated with level of formal education, and that students generally make statistically reliable gains in principled moral reasoning during college" (p. 291). However, as Nucci and Pascarella also point out, it is possible that a great deal of this growth is due to maturation, not college attendance, and that the attribution of gains to any particular type of curricular or extracurricular experience is even more problematic.

The Perry scheme describing the intellectual and ethical development of college students is similar to Kohlberg's, but focuses on students' orientations toward knowledge and authority (Perry, 1981). Research on the Perry scheme has also resulted in several instruments. One, the Reflective Judgment Interview asks students to consider four dilemmas (Brabeck, 1983). Although it has been used in a good many research projects, its interrater reliabilities are sometimes low; in general its psychometric status is unclear. A system for rating students' written reactions to questions about their college experiences and preferences, according to the Perry scheme, the "Measure of Epistemological Reflection" has been developed by Baxter-Magolda and Porterfield (1985). For example, students' reasons for preferring classes where students do a lot of talking or classes where students don't talk very much are rated. As the authors note, this measure is in the early stages of development.

2. On the meaning of critical thinking, Ennis (1986) provides a discussion and a model. Beyer (1985) discusses different definitions, Paul (1984) discusses the history of the concept, Glaser (1985) describes its role in society, and Sternberg (1985) discusses some theoretical perspectives. An issue of National Forum, (1985) was devoted to the meaning and the teaching of critical thinking.

The philosophical discussion about the generality of critical thinking is joined by McPeck (1981), Geertz (1983) and Rorty (1982). The research on the question of generality is most conveniently found in Chipman, Segal, and Glaser (1985), and

Segal, Chipman and Glaser (1985), particularly the chapters by Bowen, Perkins, Hayes, and Meichenbaum.

Alternative frameworks for considering critical thinking skills as part of a larger analysis of thinking can be found in the discussions of what constitute good thinking generally, (Nickerson, Perkins, and Smith 1985), the use of "frames" for thinking (Perkins 1986), and the "triarchic" approach to intelligence (Sternberg, 1986).

On the assessment of critical thinking, the analyses of Dressel and Mayhew (1954) are still valuable. Baron (1986) provides a good overall view, and Bransford, Sherwood and Sturdevant (1986) discuss the evaluation of programs. McMillan (1987) provides a trenchant analysis and summary of evidence on the particular assessment of critical thinking in higher education.

Two measures are predominant among those available for the assessment of critical thinking. The more widely used is the Watson-Glaser Critical Thinking Appraisal which assesses induction, assumption identification, deduction, "conclusion-logically-following-beyond-a-reasonable-doubt," and argument evaluation. The criticisms and the limitations of this measure are noted throughout this essay. The Cornell Critical Thinking Test by Ennis and Millman, is aimed at college students and other adults, and includes sections on induction, credibility, prediction and experimental planning, fallacies (especially equivocation), deduction, definition, and identifying assumptions. This instrument has been used much less in studies reported in the literature, so that it is more difficult to judge its utility in practice.

3. The Watson-Glaser and the Cornell measures have been evaluated by a panel of psychologists specializing in critical thinking in terms of whether they met the American Psychological Association's Standards for Educational and Psychological Tests (Modjeski and Michael, 1983). These judges varied considerably in their opinions. Both tests were faulted for their lack of investigation into possible test bias regarding ethnic, gender and other groupings, the lack of cross-validation in samples other than the original one used in development, and for not conducting studies of the stability of the tests over time. Both tests also received relatively low ratings of the description and rationale for the criteria; the majority of ratings were "meets standards minimally" and "does not meet standards."

4. Although this essay has concentrated on critical thinking as an example, similar considerations apply to the even more amorphous concept of "creativity." As a general introduction to the area, the reviews and syntheses of Gardner (1982, 1983) can be recommended. These reviews, as well as those by Baird (1976,

1985) suggest that creativity can best be approached by disciplinary area. In particular, the attempts to assess observable creativity can be done with some accuracy and utility within disciplinary areas. Examples of this approach to creativity can be found in Frederiksen and Ward (1978) and Baird and Knapp (1981).

An approach with considerable promise would be to build upon the work of Frederiksen and Ward, who developed a stable measure of the ability to formulate scientific hypotheses when subjects were asked to read a description of an experiment or field study, or study a graph or table showing the results and then to write possible explanations or hypotheses. Other tests asked students to evaluate proposals, solve methodological problems and attempt to measure constructs, all asking subjects to produce their own solutions. Responses are scored for their quality, number and originality. A machine-scored version did not seem as useful as the free response version (Ward, Frederiksen and Carlson 1980). Although primarily conducted within the discipline of psychology, their work serves as a prototype for other assessments of discipline based on creativity. Indeed Frederiksen et al (1981) later suggested similar assessments for medical education.

5. Under a contract from the Office of Educational Research and Improvement of the U.S. Department of Education, researchers at Iowa State are investigating the effects of differential coursework on the general learned abilities of college students. The assessment is the GRE/General Examination disaggregated by item-type. The subject universe is a sample of 1,600 college seniors at six institutions: Georgia State University, Ithaca College, Mills College, Stanford University, Florida A&M University, and Evergreen State College. The final report of this project is expected in the fall of 1989.

References

Association of American Colleges. Integrity in the College Curriculum: A Report to the Academic Community. Washington, D.C.: Association of American Colleges, 1985.

Baird, L.L. Using Self-Reports to Predict Student Performance. New York: College Board, 1976.

Baird, L.L. Review of Problem Solving Skills. Research Report R.R.83-16. Princeton, NJ: Educational Testing Service, 1983.

Baird, L.L. "Do Grades and Tests Predict Adult Accomplishment?" Research in Higher Education, vol. 23 (1985), pp. 3-85.

Baird, L.L. and Knapp, J.E. The Inventory of Documented Accomplishments for Graduate Admissions. Princeton, NJ: Graduate Record Examinations Board, Research Report 78- 3R, 1981.

Banta, T.W. and Schneider, J.A. "Using Faculty-Developed Exit Examinations to Evaluate Academic Programs." Journal of Higher Education, vol. 59 (1988), pp. 69-83.

Baron, J.B. "Evaluating Thinking Skills in the Classroom," in J.B. Baron and R.J. Sternberg (eds.), Teaching Thinking Skills: Theory and Practice. New York: W.H. Freeman, 1986, pp. 221-248.

Baxter-Magolda, M. and Porterfield, W.D. "A New Approach to Assess Intellectual Development on the Perry Scheme," Journal of College Student Personnel, vol. 26 (1985), pp. 343-351.

Bergquist, W.H., Gould, R.A. and Greenberg, E.M. Designing Undergraduate Education: a Systematic Guide. San Francisco: Jossey-Bass, 1981.

Beyer, B. "Critical Thinking: What is it?" Social Education, vol. 22 (1985), pp. 270-276.

Bowen, H.R. Investment in Learning. San Francisco: Jossey-Bass, 1977.

Brabeck, M. "Critical Thinking Skills and Reflective Judgment Development: Redefining the Aims of Higher Education. Journal of Applied Developmental Psychology, vol. 4 (1983), pp. 23-34.

Bransford, J.D., Sherwood, R.D., and Sturdevant, T. "Teaching Thinking and Problem Solving," in Baron and Sternberg, pp. 162-182.

Chipman, S.F., Segal, J.W., and Glaser, R. (eds.). Thinking and Learning Skills: Current Research and Open Questions. Hillsdale, NJ: Lawrence Erlbaum, 1985.

Colby, A., Kohlberg, L., Gibbs, J., Candee, D., Speicher-Dubin, B. Kauffman, K., Hewer A., and Power, C. The Measurement of Moral Judgment: A Manual and Its Results. New York: Cambridge University Press, 1982.

Cronbach, L.J. Essentials of Psychological Testing, 4th ed. New York: Harper and Row, 1984.

Cuban, L. "Policy and Research Dilemmas in the Teaching of Reasoning: Unplanned Designs." Review of Educational Research, vol. 54 (1984), pp. 655-681.

Dressel, P.W. and Mayhew, L.B. General Education: Explorations in Evaluation. Westport, CT: Greenwood Press, 1954.

Elstein, A.S., Shulman, L.S., and Sprafka, S.A. Medical Problem Solving: An Analysis of Clinical Reasoning. Cambridge, MA: Harvard University Press, 1978.

Ennis, R.H. "A Taxonomy of Critical Thinking Dispositions and Abilities." In Baron and Sternberg, pp. 9-26.

Fitzpatrick, R. and Morrison, E.J. "Performance and Product Evaluation." In R. L. Thorndike, Educational Measurement, 2nd ed. Washington, D.C.: American Council on Education, 1971, pp. 237-270.

Frederiksen, N., and Ward, W.C. "Measures for the Study of Creativity in Scientific Problem Solving." Applied Psychological Measurement, vol. 2 (1978), pp. 1-24.

Frederiksen, N., Ward, W.C., Case, S.M., Carlson, S.B., and Samph T. Development of Methods for Selection and Evaluation in Undergraduate Medical Education. (ETS RR 81-4). Princeton, NJ: Educational Testing Service, 1981.

Gardner, H. Art, Mind and Brain: A Cognitive Approach to Creativity. New York: Basic Books, 1982.

Gardner, H. Frames of Mind: The Theory of Multiple Intelligences. New York: Basic Books, 1985.

Geertz, C. Local Knowledge: Further Essays in Interpretive Anthropology. New York: Basic Books, 1983.

Glaser, R. "Education and Thinking--the Role of Knowledge." American Psychologist, vol. 39 (1984), pp. 93-104.

Glaser, E.M. "Critical Thinking: Educating for Responsible Citizenship in a Democracy." National Forum, vol. 65 (1985), pp. 24-27.

Gronlund, N.E. Constructing Achievement Tests. Englewood Cliffs, NJ: Prentice Hall, 1982.

Gronlund, N.E. Measurement and Evaluation in Teaching. New York: Mcmillan, 1985.

Kohlberg, L. "Moral Stages and Moralization: The Cognitive-Developmental Approach." In T. Lickona (ed.), Moral Development and Behavior: Theory, Research and Social Issues. New York: Holt, Rinehart & Winston, 1976, pp. 31-53.

McMillan, J.H. "Enhancing College Students' Critical Thinking: A Review of Studies." Research in Higher Education, vol. 26 (1987), pp. 3-30.

McPeck, J.E. Critical Thinking and Education. New York: St. Martin's Press, 1981.

Modjeski, R.B., and Michael, W.B. "An Evaluation by a Panel of Psychologists of the Reliability and Validity of Two Tests of Critical Thinking." Educational and Psychological Measurement, vol. 43 (1983), pp. 1187-1197.

Morrill, R. Teaching Values in College. San Francisco: Jossey-Bass, 1980.

Nickerson, R.S., Perkins, D.N., and Smith, E.E. The Teaching of Thinking. HIllsdale, NJ: Lawrence Erlbaum, 1985.

Nucci, L. and Pascarella, E.T. "The Influence of College on Moral Development." In J. Smart (ed.) Higher Education: Handbook of Theory and Research, Vol III. New York: Agathon Press, pp. 271-326.

Pascarella, E. "College Environmental Influences on Learning and Cognitive Development: A Critical Review and Synthesis." In J. Smart (ed.), Higher Education:Handbook of Theory and Research, Vol I. New York: Agathon Press, 1985, pp. 1-61.

Paul, R.W. "Critical Thinking: Fundamental for Education in a Free Society." Educational Leadership, vol. 42, no. 1 (1984), pp. 63-64.

Paul, R.W. "Dialogical Thinking: Critical Thought Essential to the Acquisition of Rational Knowledge and Passions. In Baron and Sternberg, pp. 127-148.

Perkins, D.N. "General Cognitive Skills: Why Not?" In Chipman, Segal, and Glaser, pp. 339-364.

Perkins, D.N. Knowledge as Design. Hillsdale, NJ: Lawrence Erlbaum, 1986.

Perkins, D.N. "Thinking Frames: An Integrative Perspective on Teaching Cognitive Skills." In Baron and Sternberg, pp.41-61.

Perry, W. "Cognitive and Ethical Growth: The Making of Meaning." In A.W. Chickering and Associates, The Modern American College. San Francisco: Jossey Bass, 1981, pp. 76-116.

Priestly, M. Performance Assessment in Education and Training: Alternative Techniques. Englewood Cliffs, NJ: Educational Technology Publications, 1982.

Quellmalz, E.S. "Developing Reasoning Skills." In Baron and Sternberg, pp. 86-105.

Rest, J. Revised Manual for the Defining Issues Test. Minneapolis: Moral Research Projects, 1979.

Rorty, R. Consequences of Pragmatism: Essays, 1972-1980. Minneapolis: University of Minnesota Press, 1982.

Segal, J.W., Chipman, S.F., and Glaser, R., (eds.). Thinking and Learning Skills, Volume 1: Relating Instruction to Research). Hillsdale, NJ: Lawrence Erlbaum, 1985.

Sternberg, R.J. "Teaching Critical Thinking, Part 2: Possible Solutions." Phi Delta Kappan, vol. 67 (1985), pp. 277-280.

Sternberg, R.J. Intelligence Applied. San Diego: Harcourt, Brace, Jovanovich, 1986.

Toulmin, S. Knowing and Acting: An Invitation to Philosophy. New York: Mcmillan, 1976.

Tuma, D.T., Rief, F., and Glaser, R. (eds.). Problem-Solving: Issues in Teaching and Research. Hillsdale, NJ: Lawrence Erlbaum, 1980.

Ward, W.C., Frederiksen, N. and Carlson, S.B. "Construct Validity of Free-Response and Machine Scorable Forms of a Test." Journal of Educational Measurement, vol. 17 (1980), pp. 11-30.

Warren, J.R. "Assessment at the Source: What is Assessment and Why Are We So Obsessed With It?" Liberal Education, vol. 73, no. 3 (1987), pp. 2-6.

Whitely, M.M. "Extracurricular Influences on the Moral Development of College Students." In M.L. McBee, (ed.), Rethinking College Responsibilities for Values. San Francisco: Jossey-Bass, 1980, pp. 45-50.

Assessment of Basic Skills in Mathematics

by Mark I. Appelbaum

When considering the issue of the assessment of basic skills in mathematics, it must be understood that differing audiences hold vastly different concepts of what constitutes basic skills. To the mathematics faculty of major research universities, basic skills may encompass pre-calculus through elementary functions and analytic geometry; to the faculty of computer science, basic skills may include topics in discrete mathematics; to the instructor of an intermediate level statistics course, the ability to solve two equations in two unknowns in order to accomplish a power analysis may be basic; to the faculty in a business school, the fundamental computational skills to handle the principles of accounting, interest, and finance are essential; while to the dean of the general education curriculum, the procedural and conceptual skills needed to complete the minimum required math course(s) for graduation may be the basic skills of interest.

In this essay we will begin by emphasizing only a narrow subset of what might be considered the proper domain of assessment in basic skills in mathematics, namely, the assessment of those procedural and conceptual skills, principally algorithmic, in basic computation, algebraic manipulation, and quantitative problem solving which are necessary for the successful completion of college level pre-calculus mathematics and math dependent courses. The exact skills required of the individual student will depend, of course, on such factors as the nature of the student's eventual major and the requirements of the general education curriculum. My intent is limited: to emphasize the basic technical requirements of such assessments, to describe some commonly used approaches and their deficiencies, and to consider some alternative approaches resulting from advancements in technology and the changing nature of demands for mathematical proficiency.

A General Model

Prior to examining these issues, it is necessary to have a general understanding of the goals of the assessment procedure in this basic skills area. As has been noted earlier in this volume, there are many purposes for which assessment technology can be used. With regard to our topic, three of these purposes seem to dominate: placement, exemption, and certification. While these three practices are closely related, there are some distinctions among them which should be considered.

By "placement" we mean the assignment of a student into an appropriate course of study--appropriate in the sense that the student both possesses the prerequisite skills needed to function successfully in the course or program, and that the course or program meets the student's anticipated needs in terms of the skills and knowledge which are to be gained. By "exemption" we mean the process by which it is determined that, because of previously acquired knowledge and skills, a student does not need to take one or more courses which would have been required had such a determination not been made. Finally, by "certification" we mean the process by which it is determined that the student possesses the knowledge equivalent to that which would be obtained in a course and hence the student should receive credit as if he or she had actually taken the course. Each of these uses of assessment implies the assessment of the individual student, i.e. a "diagnostic" evaluation of the student's knowledge and application skills, as well as a match of the student's knowledge and skills to the requirements of one or more courses. Since the focus of this essay is on "basic skill" level mathematics, only the first of these purposes will be considered in detail (though many of the issues apply to all three).

When considering the placement function of assessment in regard to basic skills in mathematics, we engage in a multi-stage decision procedure under which each stage has its unique demands. Generally, the first stage of this process involves determining whether the student possesses a basic level of procedural skills (i.e. can the student perform the required operations?) and conceptual skills (i.e. does the student understand the constructs underlying the computations?) which are thought to be prerequisite for the successful completion of any college level mathematics or math dependent course. These are skills which one generally expects the student will have acquired by the successful completion of those high school courses required by the institution for admission. At this stage, failure to demonstrate the required skills usually implies assignment to a remedial or developmental course or program with (at least, ideally) a reassessment of these basic skills upon completion of the remediation process.

Should the student demonstrate adequate mastery of basic skills (whether before or after remediation), then the second stage of placement occurs--the selection among alternative college level courses. This stage of placement depends upon several factors including the anticipated mathematical needs of the student (e.g. calculus for his/her major), assessment of higher level math skills (e.g. is the student prepared to be placed directly in a calculus course?), flexibility (i.e. will the placement give the student sufficient options if his or her curricular plans change?), as well as the student's sensibilities (i.e. is the student receptive to advice concerning the optimal selection of a course placement from among the alternatives?).

The fact that mathematics placement is a multistage decision process has several implications for the design of an assessment.

Among them: that a wide range of skills must be examined in the assessment battery, that extensive differences in basic ability and skill levels need to be accommodated by the process, and that a relatively large number of alternative courses must be considered as possible placements.

Some Technical Considerations

Despite the fact that there are multiple applications of assessment methodology in basic skill assessment in mathematics, certain basic requirements of good assessment hold across all of these applications. These include the need for "placement validity," i.e. the requirement that the content of the assessment instrument match the skills necessary for successful performance in the courses into which a student might be placed (or in the case of placement into remedial and developmental programs, the detection of the absence of certain basic skills); reliability and related psychometric properties; assurance of a lack of bias (particularly gender and racial bias); and criterion (or discriminant) validity of the cutscore (i.e. the passing or certification score). Each of these aspects of "good testing practices," including technical aspects of assessment such as reliability and validity, have already been discussed by Millman.

In the assessment of basic mathematical skills, we are dealing with assessment of individuals, hence, a much higher level of technical adequacy is required than for assessment at an aggregate level (e.g. a program evaluation). Further, since the assessment is inseparable from a series of decisions about the student, these basic principles of psychometric adequacy must apply to each and every stage of the process. Because an instrument or procedure is shown to be technically adequate for the remediation decision does not mean that it is valid for the "placement into college level courses" stage. An extensive discussion of these and related issues (including strategies for setting cutscores) can be found in Hills (1971).

Current Practices in Mathematical Skills Assessment

It is probably the case that of all the types and applications of assessment methodology in higher education discussed in this volume, none pre-dates the use of assessment for the placement of students into mathematics courses. It is likely that every reader of this volume experienced some assessment of his/her mathematical skills on entrance to college. The assessment might have been based upon performance on some admissions instrument such as the quantitative sub-test of the SAT; it might have been performance on a locally developed mathematics placement test; or perhaps it might have even been based upon the first few weeks of performance in a standard freshman mathematics course; but it is nonetheless likely that some assessment of mathematical proficiency was performed.

Given the historical concern with the assessment of mathematical skills for purposes of placement, it is surprising that a greater consensus has not emerged in the field as to how such assessments might be properly constructed and implemented. Having surveyed practices at 1,269 colleges and universities, Lederman, Ryzewic, and Ribaudo (1983) report a rather varied picture of the approaches taken to placement in basic mathematics courses. Nearly 50 percent of the reporting institutions relied solely on locally developed tests for the assessment of basic mathematic skills, about 15 percent relied on standardized admissions instruments (the SAT or ACT) alone, about 15 percent used a combination of locally developed tests and a standardized admissions instrument, about 5 percent employed state developed tests, and about 5 percent utilized what is described as "Comparative Guidance and Placement" procedures. (The remaining institutions either did not describe their placement procedures in sufficient detail or utilized such approaches as high school grades and faculty referral.) On the basis of this same survey, the authors concluded that slightly over 30 percent of entering students required help in the basic skills area of mathematics (i.e. were placed into remedial or developmental courses, or their near equivalents).

But "help" does not always mean actual enrollment in pre-collegiate mathematics courses. A 1985-1986 survey conducted by the Mathematics Association of America revealed that 15% of math enrollments in four-year colleges and 47% in two-year colleges were in pre-collegiate mathematics courses through intermediate algebra (Alders, Anderson, and Loftsgaarden, 1987). These figures may be slightly understated, though, since technical and occupational mathematics courses (many of which involve pre-collegiate skills) are listed separately.

The Lederman, Ryzewic and Ribaudo survey also revealed that while nearly 90 percent of the reporting institutions required some formal assessment of basic mathematical skills at the time of entry for purposes of placing students in mathematics courses, only 31 percent of these institutions reported using a non-course based testing procedure as a method of exiting from a basic skills mathematics requirement. By far the most commonly used criterion for demonstrating basic mathematical skills competency following placement was the successful completion of whatever course or courses the student was required to take, without any consideration of the degree to which the content of the course was related to either the content of the placement test or the future mathematical skill needs of the student.

Locally Developed Placement Tests--Their Construction

There is little published literature which addresses the nature or use of locally developed instruments for college mathematics placement; however, informal discussions with a number of individuals involved in either state-wide or local placement programs indicate several features common to a large number of these systems. First, they are almost always developed by individuals or small committees drawn exclusively from math departments, and with little input from either those who are technically knowledgeable about test development and construction or those faculty whose courses depend on the results of previous mathematical instruction.

Second, there is little knowledge of, or concern for, the technical issues of test development in these efforts; i.e. formal concern with issues such as reliability, discriminant validity, etc. This lack of concern for technical aspects of test construction is often justified by statements which reflect the point of view that these issues may not be of real import in the testing of basic mathematical skills. It is thought that the content of the field is well defined, and that there is a logically compelling structure to mathematical knowledge, and therefore any content valid test of these basic skills is adequate. But the construct validity of these tests is not always adequate, particularly when word problems are involved, as the examinee's language skills play a strong role in schemes for translating problems into tractable forms (Hinsley, Hayes, and Simon, 1977). Since students' encounters with mathematics in courses outside math and science (in psychology, business administration, economics, etc.) will most likely take the form of word problems, facility in translating schemes should be assessed. But the judgment of this facility must be distinguishable from that of algorithmic processes. Existing examinations, though, do not allow for that distinction.

In a related fashion, there seems to be a belief among those most closely involved with such tests that they tend to work fairly well in terms of effecting the proper placement of students into college level and remedial mathematics courses. If there is a perceived problem with the effectiveness of such placement tests, it is principally with the assessment of lower levels of mathematical functioning. In particular, academic advisors in selective institutions believe that a combination of scores on locally developed tests, standardized admission tests and/or Advanced Placement tests provide sufficient information to place entering students at all but the lowest levels of mathematics achievement. Finally, it is generally the case that a single test, usually administered in one to one-and-a-half hours, constitutes the complete basic skills assessment.

Locally Development Instruments--Their Content

Another striking feature of these assessment instruments is that they do not reflect recent changes in mathematical knowledge or practice. These tests look surprisingly like mathematics placement tests used 30 years ago. It may not be fair to students if such assessments define "low literacy" in basic skills mathematics in terms of what it took to be a corner grocery store owner in 1958. To move our students beyond the status of novices in mathematics in 1988 is to change the reference point of skills from the multiplication of whole numbers, for example, to multiplication algorithms (Maurer, 1985), and then to concepts that help the student explain the algorithms. To test for concepts, or for the larger structures Winograd (1977) calls "frames" is not the same task as testing for students' fluency in using algorithms. The former is analogous to explaining the principles of contrastive linguistics; the latter analogous to the more mechanical aspects of translation.

But the content of each of the tests we were able to examine was limited to (a) basic arithmetic operations with heaviest emphasis upon fractions and decimals, (b) fairly simple algebraic operations--rarely any operation more complex that solving a quadratic equation, and (c) simple graphical and tabular items. None of the tests included any items assessing the types of mathematics demanded by computer science (e.g. items which represent the ability to operate in a number system other than base 10) nor were there any items which reflected the ability to use even a simple hand-held calculator--indeed, in all but a very few of the tests examined, calculators were forbidden.

Further, very few of the tests required the application of heuristics in problem solving items, and of those which did, less than ten percent of the items were of the problem solving variety. Oddly, we do not seem to ask questions about why equations are transformed or questions that require students to classify types of equations. We give them equations to solve, but the tests do not tell us whether students understand what they have solved, hence whether they are really ready for college-level mathematics (Davis and Henkin, 1979).

In sum, most of the tests consisted of a collection of thirty to fifty items (presented in a multiple choice format) which assessed the student's ability to perform a variety of fairly low level computational operations, to engage in simple algebraic manipulations, and to extract information of a quantitative nature from graphical or tabular arrays. There was virtually no examination of quantitative problem solving strategies, and because the tests cannot probe beneath the response to a multiple-choice item, they provide little information of diagnostic value to an individual designing a remedial program. Nor did any of these exams touch on special topics which are required in math based courses taught outside of mathematics departments (e.g. the rules of summation algebra).

In fairness to the developers of these tests, it should be noted that in most four-year colleges, mathematical instruction still follows a fairly traditional sequence of pre-calculus, calculus (with perhaps a variety of introductory calculus courses stressing particular fields or applications), and then advanced courses which aim at specialized topics such as discrete mathematics, probability theory, analysis, etc. Despite calls for reform in mathematical education (Conference Board of the Mathematical Sciences,1983; Freudenthal, 1983; Raizen and Jones, 1985), problem solving is still seen as application, and the specific mathematical tools needed in applied fields are still taught in terms of isolated applications. In general, there has been rather little change in the philosophy of mathematics education at the college or university level. Given the increased importance of discrete mathematics resulting from the spread of computer technology to all fields (Ralston, 1986), the inherited curriculum reflected in these assessments may not be relevant to the academic--let alone occupational--demands placed on the majority of students. Indeed, in two-year colleges concerned with preparing students for occupations which frequently use spreadsheets, data bases and computer graphics, they may be even less relevant.

The Use of Admission Tests for Mathematics Placement

As noted by Lederman et al., standardized tests designed to be used as admission instruments (e.g. the SAT and the ACT) constitute the second most frequently used method for placement of students into mathematics courses. This use of admissions tests provides a serious source of concern (and to some extent a dilemma) when considering basic skills assessment in mathematics. As noted by the College Board (1985) in its publication, "Guide to the Use of the Descriptive Tests of Mathematics Skills":

> Although no single model for mathematics placement exists, as a general principle, faculty members should examine the content of tests being considered for use in relation to the purposes of placement and the courses into which students will be placed. If the placement decision involves a choice between assigning a student to a regular course or to a remedial course, the test to be used should measure those skills which are needed for success in the regular course and which the remedial course is intended to develop. When the placement decision concerns admission to a higher level course, the main consideration is the degree to which the test measures skills needed for that course. (p.11)

In a similar vein, the College Board (1977) recommends that when an institution uses College Board tests for purposes of placement or awarding of credit, it should "determine the appropriateness of particular tests through consultation with faculty members familiar with their content." There appears, in fact, little evidence that the use of the SAT for mathematics placement follows these suggested guidelines.

A content analysis of a recent SAT-M subtest yielded the following counts of item types (out of the sixty items which make up the SAT-M subtest):

Arithmetic reasoning - 19 questions
Algebra - 17 questions
Geometry - 16 questions
Miscellaneous - 8 questions.

> [The Miscellaneous category included items such as logical reasoning, elementary probability, general symbol manipulation, etc. Further, the SAT provides no subscale scores so that it is not possible to assess particular areas of strength and weakness.]

While geometry items constitute nearly one-third of the test, it is rare that formal geometry is included in basic skills mathematics courses; nor is formal geometry generally considered a pre-requisite skill for much of college level mathematics. Further, none of the items on the SAT examines for pure computational ability, one of the basic weaknesses found among students placed in remedial courses.

What appears to drive the continued use of such tests and what provides the dilemma in their consideration as proper assessment instruments, is their predictive power. As but a single example, Dwinell (1985) reports that among students enrolled in a sequence of courses in the Division of Developmental Studies at a southern university, the SAT-M score and high school grade point average were the best predictors of success in mathematics courses. [One might note that general psychometric theory suggests that when the range of abilities is greater than that seen among students in developmental studies, the predictive validity should, in general, be even higher.]

Commercially Developed Tests for Mathematics Placement

As an attempt to alleviate these problems while freeing institutions from the need to develop local placement tests, a number of commercial test publishers offer basic skills tests designed for diagnostic and placement purposes. One such set of instruments is the "Descriptive Tests of Mathematics Skills," a portion of the Multiple Assessment Programs and Services (MAPS) of The College Board. This set of tests contains four subtests --Arithmetic Skills, Elementary Algebra Skills, Intermediate Algebra Skills, and Functions and Graphs together with a detailed manual, "Guide to the use of the Descriptive Tests of Mathematics Skills" (1985). The manual contains sections on interpreting test scores, use of the tests for placement, technical characteristics of the tests (including information on scaling, reliability and validity), and a very useful section on the content specifications of each of the tests. This last section allows the user to estimate the "placement validity" of the test according to College Board guidelines for any particular course--provided that information on the prerequisite skills required for the course is available.

A number of other instruments for course placement and course certification in mathematics, such as the CLEP tests and the College Basic Academic Subjects Examination (currently under development), are or may soon become available (see Appendix 2 of this volume). Most of those instruments developed prior to 1985 lack content specifications in their documentation, however, and therefore require "placement validity" studies prior to their adoption.

State-wide Programs

Until recently, the assessment of basic mathematical skills has been left to the local institution (as noted above, as recently as 1983 only 5 percent of reporting institutions employed a state-wide placement instrument). Within the past 5 years, however, a number of States have taken a much more active role in setting standards for required levels of proficiency of college students in basic skills. Notable among these programs are those of Texas (the Texas Academic Skills Program) and New Jersey (the New Jersey College Basic Skills Placement Test). Both of these programs are the result of direct actions of State legislatures with respect to publicly funded institutions of higher education, and reflect the general perception of a diminishing quality of the educational system as a whole.

The New Jersey College Basic Skills Placement Test is

> designed to measure certain basic language and mathematics skills of students entering New Jersey colleges. The primary purpose of the two mathematics sections (Computation and Elementary Algebra) is to determine whether students are prepared to begin certain college level work without a handicap in computation or elementary algebra.
>
> --Interpreting Mathematics Scores on the New Jersey College Basic Skills Placement Test, p.4.

The Texas program is described as

> . . .an instructional program designed to ensure that students attending public institutions in Texas have the basic academic skills necessary to be successful in college-level coursework. The TASP will provide advisory programs and remedial support for those students who demonstrate a need to develop the basic academic skills necessary for

> success in undergraduate degree programs. Further, TASP includes a testing component. The purpose of the test is to identify and provide diagnostic information about the basic academic skills of each student. . . . Students must take the basic academic skills test before completing their first nine semester credit hours of college coursework and must pass it before completing 60 semester credit hours, or be limited to lower division coursework until they pass.
>
> --Texas Academic Skills Program,
> Program Summary, p.1.

Both of these programs include components which are designed to assess, on a state-wide basis, minimum basic skills in mathematics and to aid in placements in "remedial" level courses. This purpose of the New Jersey test is described as

> . . .placement at levels at and below the first-level college courses. It is designed to be relatively easy for well prepared students and to discriminate among underprepared students, thus affording colleges the needed range of scores to facilitate placement at several remedial levels.
>
> --New Jersey College Basic Skills Placement Testing, Fall 1987, page 4.

And as part of the enabling legislation (Texas H.B. No. 2182) of the Texas program one finds the following language:

> ...(c) The test instrument adopted by the board must be of a diagnostic nature and be designed to provide a comparison of the skill level of the individual student with the skill level necessary for a student to perform effectively in an undergraduate degree program.

and

> ...(f) If the test results indicate that remedial education is necessary in any area tested, the institution shall refer the student to remedial courses or other remedial programs made available by the institution.

as well as

> ...(g) A student may not enroll in any upper division course ... until the student's test results meet or exceed the minimum standards in all test scores.

By legislation, in neither of these States can scores from the testing components of the program be used as part of an admissions criterion (the tests are not taken until after the student has been accepted by the undergraduate institution).

Further, there are no provisions in the enabling legislation which prevent individual institutions from employing additional tests for the purpose of placement should the state mandated instrument provide insufficient information for all mathematics placements. This situation may pertain at selective colleges and at institutions in which a substantial number of students enroll in mathematically intensive programs such as engineering. In general, though, the content of the mathematics components of these state-wide tests are items which are included in courses at or below the level of the first high school algebra course and clearly focus on the more procedural aspects of mathematics.

Other states have taken somewhat different approaches to the issue of state-wide placement tests. One program of interest is the Ohio Early College Mathematics Placement Testing Program (EMPT). The purposes of the Ohio EMPT program are "(1) to inform high-school juniors of their present level of math proficiency and (2) to compare those levels to college entrance requirements." Over 900 Ohio high schools voluntarily participate in this program, and use test reults to advise their students who need additional preparation in math. From the perspective of colleges, while the EMPT test scores and suggested placements levels are only advisory, some institutions, including The Ohio State University, are currently using these scores for placement purposes. Recently, a Calculus Readiness Test has been added to allow for assessment at the more advanced level.

An interesting side feature of the state-wide programs, be they mandatory or voluntary, is a reported increase in the number and level of mathematics courses taken by students (New Jersey Basic Skills Council, 1986) as well as some changes in the content of these courses. Thus these programs have an impact not only upon placement and remedial work, but also on basic course taking and instruction.

Assessing the Effects of Remediation

While much of the emphasis of the assessment of basic skills in mathematics is for the purpose of placement (with particular emphasis on placement into remedial level courses), there is surprisingly little attention placed on the outcome of remedial programs. As noted by Lederman, relatively few institutions require outcomes testing of individual students to assure that once they have completed the remedial course(s) or program that they do indeed possess the skills that they were originally judged to be lacking. As the New Jersey Basic Skills Council (1986) notes:

> College-level courses should be conducted on the expectation that students possess the skills needed to succeed in the course. Therefore, placement criteria

> should be established carefully so as to allow students the opportunity to demonstrate these skills. Similarly, exit criteria from remedial programs should be developed to assure that students are entering college-level courses with the skills they need to succeed. Whatever level of skills proficiency a college determines for entrance into a college-level course should apply equally to students who are initially placed in that course and to students who come to the course by way of a remedial program.

It should be clear that in any assessment of basic skills in mathematics, the end point cannot be simply an evaluation of the initial placement of students into courses. Particularly in the case of students placed into remedial level courses, there must be an outcome assessment which examines for proficiency in those areas which are deemed to be necessary for successful completion of at least the general education level mathematics courses. Such an assessment should be based upon a parallel form of the instrument which was used for the initial placement, or at least which examines for the same general skill areas.

In a similar vein, the assessment of basic skills in mathematics should also include an assessment of those programs which purport to develop basic skills in students initially lacking them. A number of institutions have developed such assessments, but their results are not systematically reported in the literature and there appears to be no overall summary (i.e. a meta-analysis) of the cumulative results of such studies. Summaries of a number of these individual studies, however, are included within the ERIC system. One study of particular interest is that reported by Wepner (1985) in which the consequences of a remedial mathematics program developed in conjunction with the New Jersey Test of Basic Skills are examined. Among the features of Wepner's study which are of special relevance in this essay are the use of a post-test of similar design to the placement instrument and the collection of data on students' performance in later non-remedial level courses--two features which need to be carefully incorporated into the assessment.

Future Directions

Thus far we have described programs (local, commercially developed, or state-wide) which adhere to fairly traditional forms of assessment and testing. Many computer-based assessment projects have been undertaken in the last 5 to 10 years which offer the potential to augment these traditional approaches, and a brief description of a few of them may provide some ideas as to how they can be used in non-traditional ways for the assessment of basic skills in mathematics. The projects described here are on-going experiments which are a result of a collaborative effort undertaken by IBM and the College of Arts and Sciences at the University of North Carolina at Chapel Hill.

The first of these is a project directed by William Graves of the Department of Mathematics and Dean of General Education at UNC-Chapel Hill. The initial impetus of the project was to translate an existing locally developed math placement test used for all entering freshmen at the University of North Carolina at Chapel Hill to a computer administered format. The principal goal was to streamline the difficult process of administering a placement test to over 3,000 entering freshmen, scoring the test, and establishing the placements in the very brief time available during freshman orientation. Graves and his associates managed to demonstrate that the format translation was possible.

The value of this approach is that it allows for a much wider range of skills and abilities to be examined within the more-or-less fixed time available for testing, but with each student being assessed in greater detail, thereby yielding a more accurate assessment and the beginnings of an accurate diagnosis of strengths and weaknesses. Under such a system, students who show initial strengths in such areas as the ability to solve quadratic equations would not receive extensive testing in lower order algebraic operations but could be tested on items which would establish appropriate placement in, say, the first college-level calculus course. On the other side of the ability distribution, students who show initial deficits in simple algebraic operations would not be presented with a long series of advanced algebra problems which they are certain to get right only by guessing, but would be examined in more detail on items which could be used to assess their specific procedural abilities to handle such operations as fractions.

Such a system would then not only allow for a more exacting placement but could provide the instructor of a remedial (or advanced) course with more detailed information on the individual student than is currently available from simply a total score on a test consisting of fixed items. The other advantage of such a system is that the student can be repeatedly tested on items of a similar type so that the decision as to whether the student has mastered a particular operation is not based upon a one-shot assessment of that item type. Indeed, a well developed interactive system can be used to sequence the student through increasingly complex exemplars of the same basic operation in order to truly assess the degree to which the student can handle the operation in increasingly complex manifestations. The pilot testing of such a system by the Educational Testing Service is reported by Ward (1986).

A second project which has some implications for the assessment of basic skills is a component of this author's A Statistician's Tool Box, a series of programs designed to improve component of interest, Subscripts and Summation, was motivated by the observation that many students in introductory statistics courses, while having placed into relatively high level college mathematics courses (e.g., Calculus with Analytic Geometry) and having completed one or more of those courses, were still unfamiliar with such basic mathematical operations as subscripted variables and the rules and applications of summation notation--a very basic skill required in virtually all of statistics.

In the Subscripts and Summation program (currently designed only for the IBM-PC, XT, AT or their clones), the student sitting at a computer terminal is first presented with a short diagnostic test. Based upon the student's responses on that test, the program then branches to a series of tutorials, each designed to instruct the student on a particular aspect of subscripting (single or double subscripts), basic summation operations, and/or the algebra of summation. Thus, for a very specific mathematical skill, the program is able to assess a student's knowledge of these skills and immediately to develop a brief remedial tutorial. At the end of the tutorial, the student is retested to ensure mastery of the material. The tutorial is accompanied by printed text material, including a follow up problem set to test for longer-term retention of the material and application skills. The entire testing and tutorial program can usually be completed in less than 2 hours at any computer station in one of many computer laboratories located on campus. This program has been used with students ranging from sophomores to first year graduate students.

The point of these two examples is that assessment of basic skills in mathematics (particularly as it is used for the purpose of initial placement) does not have to be limited to the traditional fixed length test of limited basic skills. New technology and new understanding has opened the door to new ways of assessing basic skills in mathematics. Tests may now be developed that will allow not only for placement, but also for more meaningful diagnostics (with the hope that such detailed diagnostics would have some impact on instruction in remedial or developmental math courses). Moreover, we have learned that assessment of basic mathematical skills does not need to be a "one-shot" process conducted at the point of college entry, but can be included at any point, in any course where specific mathematical skills or knowledge is needed.

The instructional consequences of these advancements in assessment are significant: students can be trained at the point when discrete mathematical skills are needed, and when they are particularly motivated to learn those skills, rather than in a freshman year course where they can only be assured that "someday you will need to know this." There is, of course, the further implication that it is not necessary that all "remediation" be handled in a single class with fixed topics. As more handled in a single class with fixed topics. As more sophisticated tutorial programs become available, some specialized basic skills training (or review of methods once mastered but no longer recalled) can be handled as an optional component of the course in which those skills are needed.

A Final Comment

This essay has been predicated upon the view that the essential features of basic skills assessment in mathematics should flow from the existing demands for mathematical knowledge in the curriculum as it exists. This is not, however, a necessary assumption. Assessment can, and often should, be a partner in the development of curriculum. By focusing on issues of what ought to be the basic mathematical competencies of the college educated student as well as the pragmatic issues of what the student currently needs to succeed in the extant curriculum, the assessment process can become an important factor in the development of contemporary thinking about basic skills.

References

Alders, D.J., Anderson, R.D. and Loftsgaarden, D.O. Undergraduate Programs in the Mathematical and Computer Science: The 1985-1986 Survey. Washington, D.C.: The Mathematics Association of America, 1987.

College Entrance Examination Board. Guidelines on the Uses of College Board Test Scores and Related Data. Princeton, NJ: Author, 1977.

College Entrance Examination Board. Guide to the Use of the Decriptive Tests of Mathematics Skills. New York: Author, 1985

Conference Board on the Mathematical Sciences. New Goals for Mathematical Science Education. Washington, D.C.: Author, 1983.

Davis, R.B. and Henkin, L. "Aspects of Mathematics Learning That Should be the Subject of Testing." In Tyler, R.W. and White, S.H., Testing, Teaching and Learning. Washington, D.C.: National Institute of Education, 1979, pp. 60-82.

Dwinell, P. "The Validation of Variables Used in the Placement and Prediction of Academic Performance of Developmental Students." Paper presented at the annual meeting of the American Educational Research Association. Chicago, IL., April, 1985.

Freudenthal, H. "Major Problems of Mathematical Education." Educational Studies in Mathematics, vol. 12 (1981), pp. 131-150.

Hills, J.R. "Use of Measurement in Selection and Placement." In R. L. Thorndike (ed.). Educational Measurement (2nd Edition). Washington: American Council on Education, 1971, pp. 680-732.

Hinsley, D.A., Hayes, J.R. and Simon, H.A. "From Words to Equations: Meaning and Representation in Algebra Word Problems." In Just, M.A. and Carpenter, P.A. (eds.), Cognitive Processes in Comprehension. Hillsdale, NJ: Lawrence Erlbaum Assoc., 1977, pp. 89-106.

Lederman, M.J., Ryzewic, S.R., and Ribaudo, M. Assessment and Improvement of the Academic Skills of Entering Freshmen: A National Survey (Research Monograph Series Report No. 5). New York: City University of New York, Instructional Resource Center, 1983.

Maurer, S.B. "The Algorithmic Way of Life is Best." College Mathematics Journal, vol. 16 (1985), pp. 2-5.

New Jersey Basic Skills Council. Interpreting Mathematics Scores on the New Jersey College Basic Skills Placement Test. Trenton, NJ: Basic Skills Assessment Program, New Jersey Department of Higher Education, 1984.

New Jersey Basic Skills Council. Effectiveness of Remedial Programs in New Jersey Public Colleges and Universities, Fall 1983-Spring 1985. Trenton, NJ: Basic Skills Assessment Program, 1986.

Raizen, S.A. and Jones, L.V. Indicators of Precollege Education in Science and Mathematics. Washington, D.C.: National Academy Press, 1985.

Ralston, A. "Discrete Mathematics: the New Mathematics of Science." American Scientist, vol. 74 (1986), pp. 611-618.

Texas Academic Skills Program. Program Summary. Austin, Texas: Texas Higher Education Coordinating Board, 1988.

Ward, William, et al.. College Board Computerized Placement Tests: Validation of an Adaptive Testing Service. Princeton, NJ: Educational Testing Service, 1986.

Wepner, Gabriella. Assessment of Mathematics Remediation at Ramapo College of New Jersey. Paper presented at the annual meeting of the American Educational Research Association. Chicago, IL., April, 1985.

Winograd, T. "A Framework for Understanding Discourse." In Just, M. and Carpenter, P. (eds.), Cognitive Processes in Comprehension. Hillsdale, NJ: Lawrence Erlbaum Assoc., 1977, pp. 63-88.

Issues in Evaluating Measures of Basic Language Skills for Higher Education

by Stephen B. Dunbar

Basic skills in language--listening, speaking, reading, and writing--are often identified as primary goals of schooling at all levels. Just as the college professor expects thoughtful communication in written and spoken language to grow out of term paper assignments and class presentations, so too does the elementary school teacher try to nurture the development of good listening and reading habits in children at an early age. Verbal skill is basic to what it means to be an educated person; no one would deny the influence of this skill on future opportunity nor the importance of designing educational programs that enhance its development. Consequently, interest in instruments that might be used to assess verbal skills and in critical evaluation of such instruments for higher education is not surprising. This essay attempts to describe recent (though not necessarily novel) contributions to the measurement of basic language skills, particularly for the needs of higher education, and in such a way that important principles of measurement in the verbal domain are clarified. Some preliminary discussion regarding the nature of language skills provides a necessary foundation for understanding these principles.

Much has been written about the domain of language and language behavior. Linguists and grammarians argue at length about the possibly universal structure of language, the formal systems that describe internal properties of language, and the implications of their structures and systems for the way in which linguistic knowledge is represented in the minds of humans (Chomsky, 1972). Psychologists and psycholinguists have long entertained conflicting opinions about mental representations of language as knowledge of rules or merely observable patterns of verbal behavior (Skinner, 1957). Cultural anthropologists and sociologists develop theories of interpersonal communication and group dynamics based on observed interactions of individuals using language (Labov, 1972). Educational psychologists and language arts specialists do field research on both the processes and products of language behavior (Bereiter & Scardamalia, 1987). Indeed, language is such an integral aspect of human experience that it has been scrutinized across the disciplines, from points of view that are sometimes difficult to reconcile when selecting or developing instruments to measure linguistic competence or performance.

This multitude of perspectives on language notwithstanding, the information of most immediate interest regarding language skills for higher education relates to fluency in what might be termed the functional uses of language. Educators want to know how well students attend critically to spoken discourse, engage actively in that discourse, comprehend it in the written word, and express original or integrate existing ideas in their writing. Literacy in the broadest sense encompasses skills of listening, speaking, reading, and writing in such a way that comprehensive assessment must be considered something less if it doesn't sample each of these language behaviors. The question posed by this essay asks what factors should be considered in evaluating measures, whether existing or anticipated, of basic language or verbal skills. Principal issues focus on relations among language skills and the measures that tap them, the specificity of purpose in the assessment, the psychometric standards necessary to support the intended use of results, and the technical feasibility of a given approach.

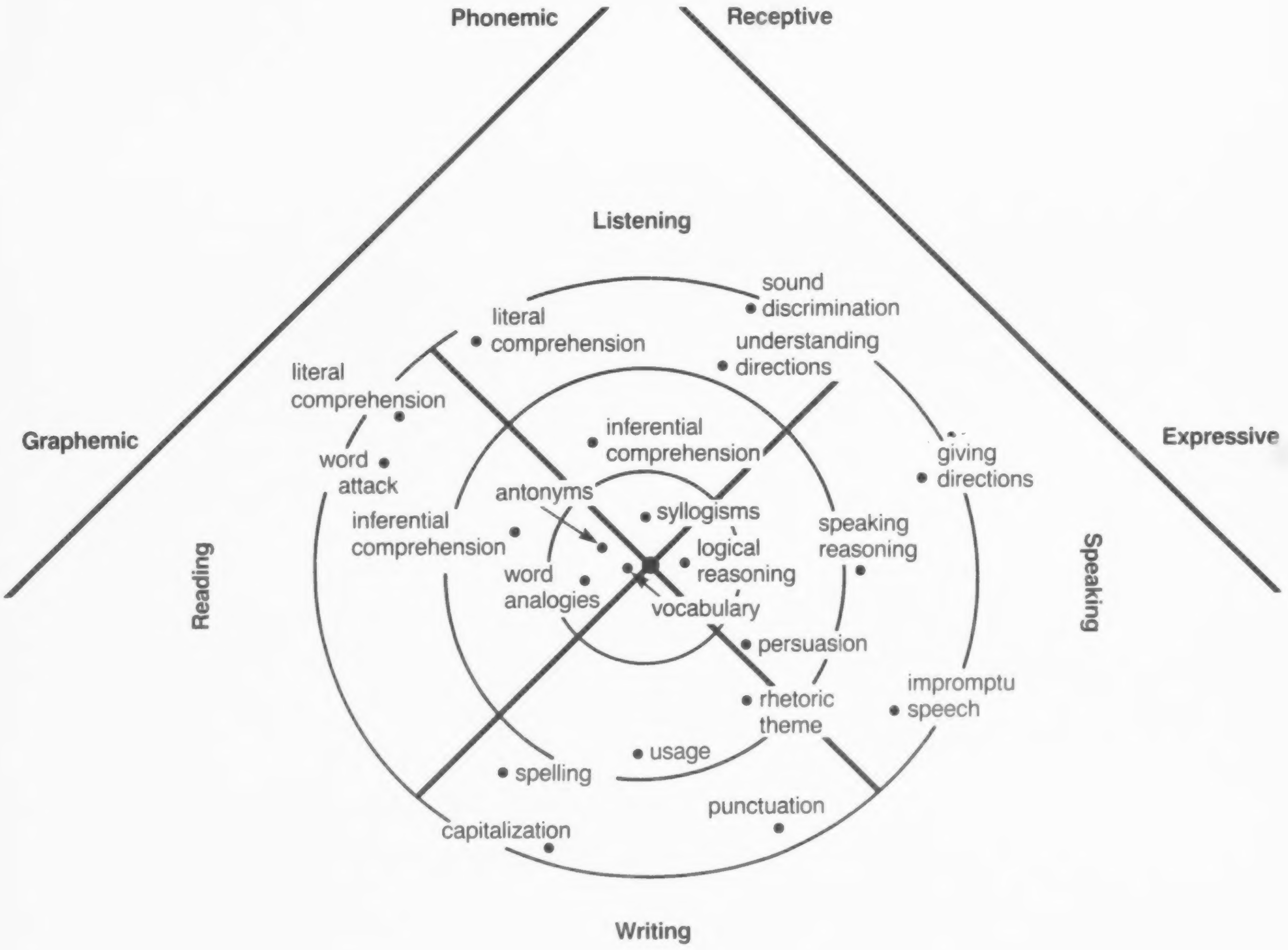
Relations Among Language Skills/Tasks

It takes no specialist in psycholinguistics to recognize that the skills discussed here are intimately related to one another. With a focus on both skills and instruments, however, developing a conceptual framework for how they might be expected to interact is important for understanding and evaluation. One such frame of reference, originally developed by Guttman (1954) as a vehicle for understanding correlations among all kinds of mental abilities tests, is the notion of a "radex" structure for psychological tests. Figure 1 presents an adaptation of the radex for understanding tests of verbal or language-related skills. Note that the description of language tests given here is primarily heuristic in that the structure discussed below has not been empirically determined. My intent is to characterize relationships among language tests that have consequences for the manner in which a user would interpret scores.

In Guttman's radex, tests are pictorially represented in a group of concentric circles, with tests located geometrically in such a way that those measuring highly related skills, abilities, or achievements will be in close proximity. Presumably, the radex structure of tests in a given content domain is validated by the identification of aspects--"facets" in Guttman's later and now the more common terminology--of test design that can explain the location of tests in the configuration. Indeed, Guttman originally proposed the radex as a conceptual model that could be used in test construction, in other words, as a general way of describing constructs of ability and achievement that would drive the development of test items (Guttman, 1970). Think of it here as a way of understanding relations among tests or tasks that could be used to drive either selection or construction of assessment procedures for language skills in higher education.

For the four language skills discussed, imagine the type of skill measured as one aspect of a test influencing its position in the diagram and interrelationships with other tests, and level

Figure 1. Adaptation of Guttman's Radex to the Domain of Verbal Tests


Phonemic
Receptive
Listening
Graphemic
Expressive
Reading
Speaking
Writing
sound discrimination
literal comprehension
understanding directions
literal comprehension
word attack
inferential comprehension
giving directions
antonyms
syllogisms
inferential comprehension
logical reasoning
speaking reasoning
word analogies
vocabulary
persuasion
impromptu speech
rhetoric theme
usage
spelling
punctuation
capitalization

of complexity in a particular skill as another. Level of complexity distinguishes measures that are easily influenced by specific instruction (e.g. knowing that commas separate non-restrictive relative clauses from the main part of a sentence) from those that are closer to being indicators of general intellectual ability. If research on tests of general intelligence (cf. Snow & Lohman, 1984) were to map directly onto the verbal domain considered here in isolation, an arrangement of tests as depicted in the figure would be expected. The concentric regions around the center are divided into wedges reflecting the four language skills. On the one hand, these skills are so arranged because they can be differentiated on the basis of their being receptive (listening, reading) or expressive (speaking, writing) uses of language. On the other hand, they are either graphemic (reading, writing) or phonemic (listening, speaking) in terms of the fundamental units of language they employ (sequences of letters or sounds, respectively). Contiguous skills are thus similar with respect to either a psychological or linguistic dimension.

Measures near the center of the radex--word analogies, logical reasoning tasks, or other integrated reasoning tasks--are at once skills most highly related to each other and least sensitive to specific instructional programs designed to promote them. Call them perhaps global outcomes of higher education that are implicit in the college experience but not acquired as a result of specific instructional interventions (like those captured in the general education objectives discussed by Centra and Baird in this volume). As one moves away from the center of the radex, the tests depicted increasingly reflect the application of specific knowledge or behavior that might be directly affected by educational opportunities, curriculum structure, or quality of teaching. Writing a book review, listening to and answering questions about a lecture on 20th century labor history in the United States, or delivering a speech that explains the steps involved in developing black-and-white photographs are examples of verbal tasks that would be located away from the center of the radex, especially if scores derived from them were aspects of common learned experiences in the population of interest. Also found in the outer regions would be other measures of general achievement in the verbal domain often used for placement purposes in higher education, such as the objective portions of the New Jersey Basic Skills Test (NJBST)[1] and Florida's College-Level Academic Skills Test (CLAST).

What is important about this framework for present purposes is not its intrinsic value in providing a mechanism or theory for psychologists to explain expected relationships among cognitive tasks involving verbal skills, but rather its instrumental value as a means of characterizing attributes of measurement techniques that might either be selected or developed for use in a college's assessment program. The radex places tasks measuring important aspects of linguistic competence and performance on a number of continua. The tasks hence differ not so much because they measure discrete language skills such as reading and writing, nor because they either are or are not immediately affected by instruction. Instead, they are viewed as relatively similar or dissimilar because of these attributes.

Illustrative Cases

To render this discussion of relations among language tests less abstract, let us consider examples of specific language assessment tools in the context of the radex structure. Purely for purposes of illustration, consider three instruments of varying formats, degrees of development, and completeness of standardization: (1) the Usage subscale of the Descriptive Tests of Language Skills (DTLS; Educational Testing Service, 1985); (2) the Reasoning subscale of the ACT College Outcomes Measurement Project (COMP; Forrest & Steele, 1982); and (3) the Communication Competency Assessment Instrument (CCAI; Rubin, 1982). The former two are components of nationally standardized examinations that measure verbal skills. The latter is included as an example of a locally developed instrument. Attributes of these instruments would suggest quite different locations in the radex structure, meaning quite different uses and interpretations of scores from them.

The DTLS were developed by the College Board to provide information to colleges about a variety of language skills, to be used generally for placement and diagnosis of individual strengths and weaknesses. They have been subjected to quite rigorous reliability and validity studies (cf. Weiss and Jackson, 1983). The multiple choice format in the Usage test presents examinees with a sentence containing four underlined phrases or clauses and asks them to identify the one in error. Cluster scores based on items measuring understanding of standard usage for pronouns, modifiers, diction and idioms, and verbs are also reported. On the average, examinees can be expected to correctly answer about two-thirds of the items on this and other tests in the battery (Weiss and Jackson, 1983), so that a student who had achieved a high degree of proficiency in the conventions of standard written English would be expected to perform exceptionally well on this test. The DTLS are the kinds of objective measures of language skills that would likely show gains as a result of remedial instruction in English grammar and usage, making them appropriate candidates for detailed evaluation of individual student progress or as a basis for directing students to individualized instructional programs. Place it in the writing wedge, but on the periphery of the radex.

By way of contrast, the DTLS can be understood as a comprehensive battery of objective language tests, whose subtests of reading comprehension, logical relationships, vocabulary, usage, and sentence structure would be scattered throughout the reading and writing sectors of the radex. Portions of the battery are represented to one degree or another in institutional placement exams such as the NJBST, rising-junior exams such as the CLAST, and other standardized language tests like the College Board's Test of Standard Written English (TSWE; Educational Testing Service, 1980).

The features that distinguish many of these tests relate less to format and content than they do to scope of coverage and difficulty. The TSWE, for example, tests only English usage in a format nearly identical to that used in the DTLS and the NJBST, but tests these skills at a higher level of difficulty. It thereby becomes, by design, a test that spreads out examinees along a wider ability continuum and predicts performance accurately enough for its use in competitive selection situations. The NJBST and CLAST include writing samples as subtests and report total scores based on both objective and non-objective parts of the exam. In a particular situation, an administrator choosing an instrument would first consider the kinds of language skills important for a given assessment. The purposes of placement in remedial courses might be better served by a comprehensive diagnostic instrument like the DTLS, whereas the certification of writing skill might be based on a combination of scores on objective and essay exams.

The Reasoning subscale of the ACT-COMP (the general exam, not the objective portion in more common use) is a composite of scores based on two types of ratings of responses to tasks requiring that examinees either write or speak on a given topic. Steele (1986) describes the measures of reasoning skills embedded in the writing and speaking components of the battery in terms of the "ability to solve social, scientific-technological, and artistic problems and to clarify social, scientific, and artistic values" (p. 4). Raters evaluate an essay and a speech provided by the examinee for the quality of reasoning demonstrated. Given sufficiently reliable ratings that successfully differentiate reasoning ability from the more formal linguistic characteristics of the essay or speech, the COMP Reasoning scale would be expected to lie closer to the center of the radex, somewhere near the boundary of the writing and speaking wedges in that the particular skills demanded by the tasks are at once complex and less likely to have been explicitly dealt with in college curricula. Generally, one would expect reliable measures of so-called higher-order skills to be near the center of the radex.[2]

The CCAI is a locally developed battery of measures of speaking and listening skills that are specifically taught in many communication studies or public speaking courses on college campuses. This instrument consists of a three-minute extemporaneous speech on a topic of choice, evaluated by raters in terms of volume, rate, clarity, gestures and the like, and oral responses to questions based on a video-taped class lecture. Rubin (1982) discusses the development of the CCAI as a response to the inappropriateness of readily available assessments in the context of the goals established at her institution regarding basic skills in listening and speaking. Again, assuming that reliable scores on this instrument are obtained by a local implementation, CCAI would be expected to lie in the speaking wedge of the radex, farther from the center than the COMP Reasoning in that its goals and scoring protocols are by definition more closely tailored to curriculum guidelines. Because the CCAI consists of both listening and speaking components, its location would in principle help define the boundary between the listening and speaking wedges of the radex.

The above examples are intended to illustrate how the radex might be used to conceptualize the vast array of instruments measuring language skills described in measurement reviews such as Buros (1986) or Stiggins (1981). Given this orientation to language testing, however, certain meaningful questions do not have transparent answers. How, for example, does one determine what characteristics a good test of listening or speaking ability possesses? When is a test too closely tied to patterns of elective coursework to be effective as a general tool for institutional research and evaluation with respect to all students? Or, what specific instruments measure such general aspects of mental ability through written or spoken language that they are not good choices for demonstrating gains to be attributed to an innovative curriculum in, say, the liberal arts?

Although Guttman’s radex provides an effective basis for differentiating verbal tasks, it does not provide simple answers to these difficult questions in higher education assessment. What consititutes a good test of a language skill very much depends on the kind of inference that is made from the test score, the technical quality of the instrument or procedure that produces the test score, and the administrative constraints that surround a contemplated assessment. These are additional issues that must guide serious discussion of measurement in the verbal domain.

Specificity of Purpose

The purposes and goals that individual institutions establish for assessment programs are likely to be as varied as the institutions themselves and the faculty and students that comprise them. Some will embrace the notion that individual academic majors have legitimate educational objectives related to verbal skill that are not necessarily universal across the campus and decide that the value or outcome of the educational experience is best determined within the major. Others will view assessment principally as a vehicle for accurate placement of incoming freshmen or transfer students, as opposed to a means of determining the outcomes of the college experience and will accordingly seek general indicators of knowledge and achievement appropriate for all students. No single essay can hope to come to terms with every potential purpose for measuring verbal abilities. However, recognizing a purpose and challenging its specificity are important considerations in putting the radex model described above to good use.

A vast repertoire of tasks that can conceivably be used to measure facility in the functional uses of language exists. Standardized achievement and proficiency tests or measures of developed language abilities with norm-referenced interpretations dominate the lists of available instruments that will be perused by committees charged with designing assessment programs (see the Resource Supplement at the end of this volume). They are typically measures that have grown out of differential psychology and been designed to maximize individual differences, leading to stable, trustworthy rankings of students in terms of general ability or knowledge. Such measures represent half of a dichotomy explicated by Lindquist (1936), who juxtaposed them with instruments better suited to "discover specific weaknesses, errors, or gaps in a student's achievement" (p. 20). In writing more recently about this other half of the test builder's dichotomy, Linn (1980) observed that "the goal of measuring achievement is much more elusive than the goal of differentiating among individuals" (p. 84). It is a goal that defies subject matter specialists to describe exhaustively or measurement specialists to model statistically, but it is nevertheless a goal that requires greater specificity of purpose than does the goal of ranking students for accurate prediction of a relevant external criterion. It is, moreover, the goal that embodies the aspirations of many in higher education who view assessment as a means to a pedagogical end (Adelman, 1986).

The Texas Experiment

Recent legislation passed by the State of Texas (House Bill No. 2182) is a useful case in point, not for the particular guidelines established for the basic skills assessment component of this far-reaching bill, nor for the desirability of mandated assessment programs in higher education, but rather for the directness of the charge to the Texas Higher Education Coordinating Board (THECB), assigned the task of developing and implementing the program. It is also instructive insofar as language skills represent two-thirds of the effort. The legislation specifies that the instruments to be used "must be of a diagnostic nature and be designed to provide a comparison of the skill level of the individual student with the skill level necessary to perform effectively in an undergraduate degree program" (House Bill No. 2182, pp.1-2). The target areas in the resulting Texas Academic Skills Program (TASP) are reading, writing, and mathematics, and the program as a whole represents "an extension of existing requirements for students entering teacher preparation programs" (THECB, 1987), which specify that prospective students pass a basic skills test as a condition for admission to the education major (not as a condition for university admission, however).

It is too early in the development of TASP to examine critically the quality of various components in the Texas program. The impetus of the legislation is clearly remediation in the basic skills, and the most extensive components in terms of development and execution focus on expanded educational opportunities through remedial instruction, not on testing. However, the tests that will be developed are clearly positioned as the catalysts for change in public higher education and as such must be able to withstand public and professional scrutiny.

Officials of THECB are now involved in an elaborate scheme of instrument development, which in the two language-related content areas involves instruments of quite different formats. A test of reading proficiency and a writing sample are being designed with guidelines based on results of faculty surveys, an advisory committee of some thirty educators from around the State, and regional and minority review panels. On visual inspection the reading test will likely resemble the kind of content-referenced reading comprehension achievement test that is typical in statewide assessments, but will have to be sufficiently detailed in its design to provide the kinds of "diagnostic and prescriptive" score reports stipulated by the legislation. Because the target population for this program and the level of proficiency required are somewhat unique--implicit in the legislation is a view that commercially available tests of reading and writing will not adequately address the concerns for remediation--one would anticipate that the instruments growing out of it will reflect a more detailed definition of the components of the reading process (cf. Curtis and Glaser, 1983) than is represented, for example, by the three cluster scores of the DTLS Reading test (understanding main ideas, understanding direct statements, drawing inferences). The board may well sample from various levels of a scheme like Guttman's radex structure by the time it defines diagnostic cluster scores accomodating the views of all concerned with the design of the reading test.

The development that will be particularly interesting to watch in the Texas case concerns the writing sample, in that the same types of score reports, those that identify strengths and weaknesses and suggest directions for remediation, are called for in writing as well. As described in the essay on direct writing assessment in this volume, language tests that base scores on the judgments of expert raters require careful monitoring of topics and scoring protocols to ensure acceptable levels of reliability and validity. Research in direct assessment of abilities in writing and speaking generally shows that global or so-called holistic judgments of quality are more reliable than particularized judgments of the kind that would appear to be required by the Texas legislation. The diagnostic requirements of the Texas assessment program create new challenges in the development and large-scale implementation of scoring protocols appropriate for college-level writing. Such challenges will have to be met in order to produce scores from writing samples with the degree of technical quality expected for the individual assessments of strengths and weaknesses that the Texas bill stipulates, and to the degree demanded by the "high stakes" use to which results are to be put.

Issues of Technical Quality

Millman's essay in this volume discusses the technical standards for the quality of assessment procedures used by specialists to evaluate instruments. They needn't be reviewed again here. However, two issues in evaluating instruments and that merit special attention come to mind in the particular context of basic language skills.

The first can be understood on consideration of the format implied by many of the tasks noted around the periphery of the radex structure. It concerns the precision with which a given format can be expected to measure a given language skill. Many of the assessment procedures used to evaluate curricular goals in higher education, exemplified by TASP, entail direct observations of verbal behavior, and ratings of quality with respect to important aspects of performance. The non-objective portions of the ACT-COMP use such rating procedures, as do many other innovative approaches to assessment in higher education at institutions like Alverno College and the University of Tennessee-Knoxville (cf. Loacker, Cromwell and O'Brien, 1986; Banta, 1985). Further, Banta and Fisher (1987) made a strong case for multifaceted approaches to assessment of general goals in higher education, arguing that "when the objectives for a general-education curriculum are compared with the content of the commercial tests available, it is apparent that none of the tests measures more than half of the broad understanding most faculty members believe general education should impart" (p. 44).

A call for verbal tasks with a high degree of face validity and a close correspondence to local definitions and implementations of general education objectives deserves praise. It is a call that resonates throughout Lindquist (1951), in which he states that "the only perfectly [emphasis added] valid measure of attainment of an educational objective would be one based on direct observation of the natural behavior of the individuals involved, or of the products of that behavior" (p. 142).

However, Lindquist (1951) as well as many others since have argued that when the questions turn to the psychometric characteristics of the resulting instruments and adherence to professionally recognized standards of technical quality in development and execution (e.g. AERA, APA, NCME, 1985), due regard must be given to the complexity of using observational data for assessment. Just because a task is more difficult, however, does not imply that it should be abandoned in favor of a more readily available or efficiently administered approach. Theories of measurement that lead to workable definitions of reliability for paper-and-pencil tests exist because of similar demands from educators relating to purposes of selection and placement. The call for direct and systematic observations will no doubt echo until measurement theorists develop efficient mechanisms for characterizing their technical properties. Braun (1986) describes recent attempts to increase the efficiency of procedures for performances ratings used with essays and production-based measures, and both my essay below and Adelman's concluding essay in this volume elaborate on these issues.

Convergent and Discriminant Validity

The second technical issue hinges on quantitative relationships among measures of ostensibly different language skills and is captured in the technical language of the measurement specialist by the terms convergent and discriminant validity. It is quite possible for scores on a reading test, for example, to be so highly correlated with scores on a listening test, that the two instruments fail to differentiate the skills of reading and listening in such a way for them to have diagnostic value. Such scores would fail to have discriminant validity. Campbell and Fiske (1959) introduced these terms to describe construct validity of inferences based on test scores. Briefly, tests of the same ability should be more highly related than tests of different abilities. Further, tests of nominally different abilities should yield scores that are sufficiently distinct to allow dependable diagnosis and recommendations for remediation. Although evidence for convergent and discriminant validity usually takes the form of correlation coefficients among scores for representative samples of examinees, the concepts are important for descriptions of test content as well.

Instruments used to measure certain language skills have been criticized for their lack of discriminant validity (Palmer, Groot and Trosper, 1981). Mead (1986), for example, discusses the fact that many existing language tests are related to general mental ability to such a great extent that they are inappropriately used to discriminate listening ability from reading ability within the individual. The radex structure for language tests predicts this to be the case for certain types of measures, but not for others. The domain definitions and construction practices that guide test development have an obvious impact as well. For example, close examination of some commercially available listening tests would suggest that a common strategy in the development of some measures of listening is the presentation of passages of written text traditionally used for reading comprehension tests orally, with follow up questions (Powers, 1984). Calling the resulting measurement one of listening comprehension begs the question of discriminant validity.

The same practice has been used in the modes of speaking and writing, namely, using topics and scoring criteria developed for collecting writing samples as a basis for oral presentations, and vice-versa. While it is certainly possible that substitutions of this sort could, at least in principle, produce effective measures of ability in either mode of presentation, content specialists would no doubt question the strategy as general practice. Shaughnessy (1977) notes this contrast between speaking and writing situations by saying of the latter that no "open bargaining can go on in the writing situation [for] the writer cannot keep an eye on his reader nor depend upon anything except words on a page to get him his due of attention" (p. 12). Measurement specialists would in turn question the diagnostic value of scores that result from a strategy that casually substituted writing situations for speaking situations. The purpose of the radex structure was to show that such practices might indeed lead to measures that would fail to distinguish the level of processing in listening from that in reading, or the effectiveness of communication in speaking from that in writing.

Some Institutional Concerns

The assessment concerns of the institution go well beyond the relatively formal description of the domain of language tests offered in this essay. Language behavior is necessarily cultural behavior. Over the years, educators have become increasingly sensitive to understanding diversity in language use in addition to uniformity of standards. Any assessment presupposes a standard, but exists in a community of linguistic diversity. Shaughnessy's (1977) perspective on standards of correctness offers an orientation to measurement that deserves careful consideration.[3]

Linguistic diversity has important implications for settings in which special populations have high representation in a student body. Speakers of a non-standard English vernacular and speakers of English as a second language may experience unique problems taking tests. The APA/AERA/NCME Test Standards recognize that subject-matter tests in the major, for example, or tests of general educational development can instead end up being language proficiency tests for such students. When measures are used for placement and proficiency testing in a content area, the potential for adverse impact on special populations may increase as a result of linguistic factors that are extraneous to the skills being evaluated. Although every dean would like to have specific guidelines to follow under these circumstances, the matter of linguistic diversity is one to be weighed and considered in individual cases. It should be prominent in the minds of committees charged with designing an assessment.

Most institution-wide assessments admit to a multiplicity of purposes. To ask that a language assessment merely identify a group to be targeted for remedial instruction is usually not enough. A relatively easy battery (like the DTLS or an instrument aimed at minimum competencies) provides little information regarding students that might be identified for accelerated programs or placed in upper division courses because of high levels of achievement. Multiple purposes for assessment often imply that multiple levels of difficulty in the instruments are needed. Functional-level testing (i.e. using a flexible battery of instruments that can be more closely tailored to an individual's level of ability) represents a complex but workable solution to an institution's interest in maximizing the information gained from a large scale assessment when ability levels of examinees are heterogeneous. Grandy's essay on computer-interactive testing in this volume provides a useful perspective on tailored testing.

Conclusion

Much of the foregoing discussion has concerned principles of measurement in the verbal domain that could be used to characterize properties of and relations among relevant instruments and procedures. Although the technical language of the measurement theorist has been avoided for the most part, the formal nature of their concerns has not. The underpinnings of both the radex model and its implications for the selection and development of procedures are highly technical. The precise locations of tests and other measurement devices in the radex structure, for instance, are determined by quantitative methods that are no less complex than the judgmental methods that might be used to evaluate specific content vis-a-vis the radex on the conceptual level emphasized in this essay.

Technical concerns, however, cannot be separated from local conditions that surround assessment at the institutional level. Put another way, the properties of assessment procedures will depend on factors such as the educational experiences of students at an institution and the commitment of the faculty and other key players to both the means and ends of assessment. Verbal tasks in general will be simple or complex to the extent that the educational experiences of students consistently involve active use of language as a vehicle for higher learning in the disciplines. Expectations for performance, including the kinds of judgments of threshold levels required by mandated programs like the one in Texas, cannot be properly understood without proper scrutiny of the efforts made locally to promote the skills measured in the assessment.

End Notes

1. The objective portion of the New Jersey Basic Skills Test includes a subtest called Sentence Sense, whose format is shared by many objective tests of English usage. About half of this subtest is a traditional "find the error" test. A second part requires that students rephrase sentences mentally, as they would do in editing their own written work, and to identify the most appropriate rephrasing. The second task is designed to simulate, in a multiple choice format, production skills used in editing and revising written work. In theory, this test would probably by located closer to the boundary between writing and speaking, on the one hand, and closer to other production-based measures of writing, on the other, than would a pure error recognition test.

2. A distinctive feature of measures near the center of the radex is that they are not especially sensitive to explicit instruction. Critical thinking tests, for example, tap aspects of verbal knowledge or skill that are dealt with tacitly in most college curricula and measure relatively stable characteristics of the individual. Such tests are not likely to show dramatic gains as a result of instruction. The average gain of about eight raw-score points on ACT-COMP between the freshman and senior years reported by Curry and Hager (1987) illustrates the dilemma faced by an institution trying to demonstrate gain with instruments near the center of the radex.

3. Language tests are especially prone to charges of ethnocentrism in that "standard" English usage is by definition the standard of the dominant group in a society. To defend language tests against such charges requires that one value the diversity from which they stem and to promote the standard as yet another example of linguistic diversity. Shaughnessy (1977) argues that "a person who does not control the dominant code of literacy in a society ... is likely to be pitched against more obstacles than are apparent to those who have already mastered the code. From such a vantage point ... one knows that errors matter ... " (p.13). But in her view one also knows that errors have their own logic, a logic that good teachers (and good test builders) should capitalize on for meaningful diagnosis and instruction.

References

Adelman, C. "To Imagine an Adverb: Concluding Notes to Adversaries and Enthusiasts." In C. Adelman (ed.), Assessment in Higher Education: Issues and Contexts. Washington, D.C.: U.S. Department of Education, 1986, pp. 73-82.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education Standards for Educational and Psychological Testing. Washington, D.C.: American Psychological Association, 1985.

Bereiter, C. and Scardamalia, M. The Psychology of Written Composition. Hillsdale, NJ: Lawrence Erlbaum, 1987.

Banta, T.W. "Use of Outcomes Information at the University of Tennessee, Knoxville." In P.T. Ewell (ed.), Assessing Educational Outcomes. San Francisco: Jossey-Bass, 1985, pp. 19-32.

Banta, T.W. and Fisher, H.S. "Measuring How Much Students Have Learned Entails Much More than Simply Testing Them." Chronicle of Higher Education, March 4, 1987, pp. 44-45.

Braun, H.I. Calibration of Essay Readers Final Report. ETS Program Statistics Research Technical Report No. 86-68. Princeton, NJ: Educational Testing Service, 1986.

Buros, O. (ed.). The Ninth Mental Measurements Yearbook. Highland Park, NJ: Gryphon Press, 1986.

Campbell, D.T. and Fiske, D.W. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." Psychological Bulletin, vol. 56 (1959), pp. 81-105.

Chomsky, N. Language and Mind. New York: Harcourt, Brace, Jovanovich, 1972.

Curry, W. and Hager, E. "Assessing General Education: Trenton State College." In Halpern, D.F. (ed.), Student Outcomes Assessment: What Institutions Stand to Gain. San Francisco: Jossey-Bass, 1987, pp. 57-65.

Curtis, M.E. and Glaser, R. "Reading Theory and the Assessment of Reading Achievement." Journal of Educational Measurement, vol. 20 (1983), pp. 133-147.

Educational Testing Service. Test of Standard Written English. Princeton, NJ: Educational Testing Service, 1980.

Educational Testing Service. Descriptive Tests of Language Skills. Princeton, NJ: Educational Testing Service, 1985.

Forrest, A. and Steele, J.M. Defining and Measuring General Education Knowledge and Skills. Iowa City, IA: American College Testing Program, 1982.

Guttman, L. "A New Approach to Factor Analysis: the Radex." In P. F. Lazarsfeld (ed.), Mathematical Thinking in the Social Sciences. Glencoe, IL: The Free Press, 1954, pp. 216-257.

Guttman, L. "Integration of Test Design and Analysis." In Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service, 1970, pp. 53-65.

House Bill No. 2182. Austin: Texas State House of Representatives, 1987.

Labov, W. Language in the Inner City: Studies in the Black English Vernacular. Philadelphia: University of Pennsylvania Press, 1972.

Lindquist, E.F. "The Theory of Test Construction." In H.E. Hawkes, E.F. Lindquist, & C.R. Mann (eds.), The Construction and Use of Achievement Examinations. Boston: Houghton-Mifflin, 1936, pp. 17-106.

Lindquist, E.F. "Preliminary Considerations in Objective Test Construction." In E.F. Lindquist (ed.), Educational Measurement. Washington, D.C.: American Council on Education, 1951, pp. 119-158.

Linn, R.L. "Test Design and Analysis for Measurement of Educational Achievement." New Directions for Testing and Measurement, vol. 5 (1980), pp. 81-92.

Loacker, G. Cromwell, L, and O'Brien, K. "Assessment in Higher Education: to Serve the Learner." In C. Adelman (ed.), Assessment in Higher Education: Issues and Contexts. Washington, D.C.: U.S. Department of Education, 1986, pp.47-62.

Mead, N.A. "Listening and Speaking Skills Assessment." In R. Berk (ed.), Performance Assessment: Issues and Methods Baltimore: The Johns Hopkins University Press, 1986, pp. 509-521.

Palmer, A.S., Groot, P.J.M., and Trosper, G.A. (eds.). The Construct Validation of Tests of Communicative Competence. Washington, D.C.: Assoc. of Teachers of English to Speakers of Other Languages, 1981.

Powers, D.E. Considerations for Developing Measures of Speaking and Listening. College Board Report No. 84-5. New York: College Entrance Examination Board, 1984.

Rubin, R.B. "Assessing Speaking and Listening Competence at the College Level: the Communication Competency Assessment Instrument." Communication Education, vol.31 (1982), pp.19-32.

Shaughnessy, M.P. Errors and Expectations: A Guide for the Teacher of Basic Writing. New York: Oxford Univ. Press, 1977.

Skinner, B. F. Verbal Behavior. New York: Appleton-Century-Crofts, 1957.

Snow, R.E. and Lohman, D.F. "Toward a Theory of Cognitive Aptitude for Learning from Instruction." Journal of Educational Psychology, vol. 76 (1984), pp. 347-376.

Steele, J.M. "Assessing Reasoning and Communicating Skills in College." Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1986.

Stiggins, R.J. A Guide to Published Tests of Writing Proficiency. Portland, OR: Northwest Regional Educational Laboratory, 1981.

Texas Higher Education Coordinating Board. Texas Academic Skills Program: Program Summary. Amherst, MA: National Evaluation Systems, 1987.

Weiss, D. and Jackson, R. The Validity of the Descriptive Tests of Language Skills: Relationships to Direct Measures of Writing Ability and with Grades in Introductory College English Courses. College Board Report 83-4. New York: College Entrance Examination Board, 1983.

Assessing General Education

by John A. Centra

The need to include a liberal or general education component in a college curriculum is seldom disputed. But the goals and content of that component vary from campus to campus and any assessment of general education by a college needs to begin with an understanding of the type of program and its desired effects. In addition to the type of program, the purpose of the assessment should also be considered. Is the institution trying to assess the effectiveness of its general education program? Will the results be used to certify an individual student's competence or achievement level? As Millman's essay has explained in some detail, program effectiveness and individual student certification or placement each call for different approaches to assessment and the instruments that might be used.

General education at American colleges and universities takes three broad forms: distribution requirements, core curricula, and skills or competency-based programs. Distribution requirements are by far the most common approach to general education (Levine, 1978; Klein and Gaff, 1982). Designed to ensure that students take a minimum number of courses in specified academic areas--typically the natural sciences, social sciences, and humanities--distribution requirements may be selected using either a prescribed or a "smorgasbord" approach (Levine, 1978). Colleges with prescribed distributions require students to take a given number of courses in each designated area, including some specified courses. Smorgasbord distribution requirements include few, if any, specified courses, although students must still select courses from some but not all, designated areas. There may be as many as eight of these designated areas so students could, for example, take all of their required courses in the arts, humanities and social sciences.

For institutions with prescribed distribution requirements, assessment is a matter of judging how well the students' level of knowledge and skills reflects the desired outcomes of both the specified courses and the courses students select from designated subject areas. Courses that students may select will frequently include those that are introductions to the disciplines or those that cover subject matter in the disciplines at more advanced levels. Lofty goals may be espoused by an institution for its general education program, but implementation of an accurate assessment comes down to sampling the content of the particular curricular offerings. It is unlikely, however, that a given college will find a standardized test that will exactly reflect the content of its prescribed or smorgasbord curriculum.

Several tests of general education content areas are produced by commercial publishers and these are described in the second part of this essay. Some of these, such as the new Academic Profile or the College-Level Examination Program (CLEP) General Examinations, were designed to assess distribution requirements in the humanities, natural sciences, and social sciences, but institutions should examine the test specifications before accepting them for program evaluation or individual certification. Other CLEP tests were designed to cover the content included in many college-level introductory courses in the academic disciplines, but these tests, along with the Advanced Placement Tests and the International Baccalaureate, are used principally to grant credit to individual students or to place them at the appropriate level within the discipline. Although they could be used to assess general education programs that emphasize introductory courses in the disciplines, such use would be atypical and relatively expensive (even assuming that a sample of undergraduates would be tested).

The second approach to general education, the core curriculum, is less common than distribution requirements, though it has experienced a rebirth in recent years. Core programs usually consist of an interdisciplinary series of courses focused on common problems or issues. The sequence of courses, which might include such titles as "World Views and Ways of Knowing," "Freedom and Power," or "The Modern Imagination," is generally required of all students.

The specific objectives of an institution's core curriculum can best be assessed by locally developed measures. Perhaps the test that comes closest to assessing the broad goals of many core curricula is the American College Testing (ACT) program's COMP est. The sub-scores included in the test (Solving Problems, Clarifying Values, Communicating, and Functioning Within Social Institutions, and others) reflect the type of interdisciplinary and functional emphases in many core sequences. The COMP test, however, would not be totally appropriate for at least one type of core curriculum: the four-year Great Books program at St. John's College or similar general education core curricula based on a specific set of texts. The COMP test does not examine the content of any particular text, although it may reflect specific cognitive skills addressed by this type of curriculum.

The third category of general education in American colleges and universities consists of competency-based programs. A handful of colleges such as Sterling College and (most notably) Alverno College offer such a curriculum, in which students are required to demonstrate attainment in as many as eight interdisciplinary areas. Assessment is an integral part of these programs, with locally designed and administered evaluations playing a key role in individual student and program assessment.

The development of general cognitive skills is a principal objective of most general education programs, although skill and knowledge attainment should not be regarded as separate and unequal ends (Marton, 1979). Indeed, both the ACT COMP test and the Academic Profile assess skills within a context or content area, on the assumption that people communicate or solve problems within some subject matter context. In Figure 1, a matrix of the content and skills areas of these two tests is displayed. Both tests address the same three content areas, although the Academic Profile uses the more traditional classification of natural sciences, social science, and humanities, and the COMP Test emphasizes use of information in each of these three areas (e.g. Using Science and Technology). The COMP Test provides three skills scores, two of which overlap to some extent the four Academic Profile skill areas, though with different titles (e.g. Communicating rather than Reading and Writing). As described below, institutions can obtain individual or group scores in each of these areas for both tests.

In sum, the three categories of general education emphasize somewhat different goals and, more critically, ways of attaining those goals. What to assess as well as when to assess are both affected by the type of program. Some core curricula and distribution requirements are completed within the first two years of a four-year baccalaureate, others continue into the upper levels. These variations should determine whether students are tested at the end of the second or fourth year. For two-year colleges, the general education program will generally be spread out over the entire two years for the associate degree and in the first year only for some technical areas. Ideally, testing should occur at the end of the general education sequence.

Before examining specific instruments and methods for assessing the outcomes of general education, it is helpful to repeat two principles from Millman's essay on global issues in assessment design. First, when the purpose of the assessment is program evaluation, student outcome measures are only one piece of the evidence. Qualitative evidence such as that derived from student and faculty interviews, and analyses of course syllabi can help determine how well the program as a whole is functioning. A well-designed program evaluation should involve the faculty and staff of a college, especially if the results are to be used for program improvement. Only then will the measures developed locally or selected from available commercial tests reflect program goals. More importantly, only then will the results stand a good chance of being used to make decisions about the program.

Second, when the purpose of the assessment is program evaluation, test scores need only have acceptable group reliability. In comparison to individual assessment, for which individual score reliability is important, fewer test items (and thus less testing time) are required for group reliable scores. Although it is not necessary to test every student for program

Figure 1. Content x Skills Matrix of the ACT Comp and the ETS Academic Profile

(Approximate Overlap Between Scales)

	Content Areas		
Skill Areas	Functioning Within Social Institutions (ACT) Social Science (ETS)	Using Science and Technology (ACT) Natural Science (ETS)	Using the Arts (ACT) Humanities (ETS)
Communicating (ACT) Reading (ETS) Writing (ETS)			
Solving Problems (ACT) Critical Thinking (ETS)			
Clarifying Values (ACT)			
Using Mathematical Data (ETS)			

evaluation, it is important to select a representative sample of students so that an accurate estimate of how all students would perform can be attained. If a test has adequate individual score reliability, its group scores will have even higher reliability values.

Commercially-Available Tests of General Education

One danger in the use of commercial tests for program assessment is that the tests may greatly influence the curriculum content and, in fact, faculty members may begin teaching specific test content rather than the general principles or domains on which the test is based, especially if the results are used for accountability purposes. Excessive testing or the use of inappropriate tests can also prove cost-inefficient.

Commercial tests to be reviewed in this section for use in program assessment or for individual certification are:

1. The ACT COMP Test, including the Composite Examination and the Objective Test.
2. The ETS Academic Profile.
3. The CLEP General Examinations.

Other tests have been used by colleges to assess general education, but problems in content validity render them generally inappropriate for such use. One example is the ACT Assessment examination, a test based on student achievement in high school and designed as a college admissions examination. Because it reflects the content of the high school curriculum rather than college-level general education programs, it is not appropriate for either program evaluation or individual certification for college course credit. Other inappropriate tests include the NTE Core Battery Test of General Knowledge and the GRE General Examination. The former was designed to certify students for teaching in elementary and secondary schools, and the two-hour battery in mathematics, social studies, literature and fine arts, and science consists principally of recognition items drawn from pre-collegiate curricula. The GRE/General Examination, on the other hand, provides measures of general learned abilities (verbal, quantitative and analytic), rather than achievement measures based on an undergraduate curriculum. Because the GRE/General correlates highly with the SAT, it is possible to assess the indirect effects of different general education curricula on improvement in students' general learned abilities, but the test (like the NTE) is certainly not appropriate for individual certification of general education attainment. Besides, as Banta and Fisher (1987) point out, none of the commercial tests "measure more than half of the broad understanding most faculty members believe general education should impart."

The ACT COMP Examination

In 1976, the American College Testing Program (ACT) organized the College Outcome Measures Project (COMP) to help colleges assess general education outcomes, and over 200 colleges have since conducted studies using COMP assessment instruments.

To determine which outcomes to assess, the COMP staff studied the literature on general education and interviewed faculty at a diverse group of institutions. Over 500 outcome statements were collected from the colleges and from two statewide agencies. Working with institutional and agency representatives, the COMP staff ultimately classified the statements into six domains, three each in the process and content areas.

The content areas and their general definitions are: (Forrest and Steele, 1982, pp. 57)

1) Functioning Within Social Institutions:

The ability to identify activities and institutions which constitute the social aspects of a culture, to understand the impact that social institutions have on individuals, and to analyze one's own and others' personal functioning within social institutions.

2) Using Science and Technology:

The ability to identify activities and products which constitute the scientific technological aspects of a culture, to understand the impact of such activities and products on individuals and the physical environment in a culture, to analyze the uses of technological products in a culture.

3) Using the Arts:

The ability to identify activities and products which constitute the artistic aspects of a culture, to understand the impact that art, in its various forms, has on individuals, and to analyze uses of works of art.

The three process areas and their definitions are:

1) Communicating:

The ability to send and receive information in a variety of modes and settings, including mathematical and graphical data, and for a variety of purposes (e.g. to inform, to understand, to persuade, and to analyze).

2) Solving Problems:

The ability to analyze a variety of problems (e.g. scientific, social, personal), to select or create solutions to problems, and to implement solutions.

3) Clarifying Values:

The ability to identify personal values, to understand how values develop, and to analyze the decisions made on the basis of personally held values.

Institutions may administer either the Objective Test or the Composite Examination. The Objective Test poses all questions in a unique multiple-choice format, while the Composite Examination includes both multiple-choice questions and those requiring written responses and judged by faculty raters. An Activity Inventory that assesses the quality and quantity of individuals' participation in various activities in each of the six outcome areas is also available to institutions.

The Composite Examination

The Composite Examination and the Objective Test are based on a two-dimensional matrix defining the six outcome areas. Based on the assumption that people do not communicate, solve problems, or clarify values without some content involved, the two-dimensional relationship refines further the six outcome areas listed in the margins of Figure 1.

The Composite Examination yields a maximum total score of 186, with sub-scores in each of the six outcome areas. The Communicating sub-score also includes scores for Speaking and Writing. Stimulus materials for student responses include written narratives, such as a letter or memo from a government agency; written narratives with numerical or graphical representations (e.g. an article on the economy); audiotapes, such as recorded music (to assess the use of art); videotape or film, such as a TV commentary on food production; and written, oral, or visual stimuli which require an oral response. The questions require students to provide their own short oral or written answers to two-thirds of the stimuli, or to select the two correct alternatives from among the four choices presented to them in the remaining one-third of the items. Fifteen simulation activities represent a variety of situations that test skills and knowledge considered important for effective functioning in adult roles. Faculty raters evaluate the oral or written responses using standardized rating scales; the other responses are machine-scored.

The Objective Test

Because the Objective Test is less expensive than the Composite Examination, less time-consuming to take (two v. four hours) and to score (machine scorable), and less complex to administer, it is more widely used than the Composite Examination. The Objective Test has at least three disadvantages, however: (1) It is not as reliable as the Composite Examination, although its group score reliability is good and it correlates very well (.80) with the Composite Examination; (2) It does not produce sub-scores in writing and speaking, although it does produce scores for the six major areas of COMP (with a total score of 240); and (3) It does not give local faculty the opportunity to read or listen to student-generated responses.

The Objective Test, like the Composite Examination, consists of fifteen simulation activities based on realistic stimulus materials. Students are asked to select the two correct responses among the four alternatives for each item. A penalty for incorrect responses results in each item being scored on a range from -2 to +2. This format allows a single stimulus to elicit responses for two scales and provides a strong penalty against guessing.

The following sample item from the Objective Test may also be used in the multiple choice portion of the Composite Examination or the stimulus material may be included in the free response portion of the Examination. To get some sense of the scales, the item is classified by the content and process area indicated in the COMP manual. While many of the items seem to be logically categorized, others could be classified in another content or process area. This content overlap would contribute to the relatively high inter-correlations among the scales (see discussion under validity below).

> Item: Respondents read a brief article from Newsweek on automobile energy use and answer six related questions (15 minutes total time). Students are told to "draw on all the knowledge and skills you have acquired from any source to identify answers." The article describes the dilemma that car makers face of having to produce smaller, less profitable cars to meet government regulations at a time when many consumers want larger cars. Information in the article enables students to answer the six questions, one of which is:
>
> "Answers to which of the following questions would determine if government fines are effective in solving an energy shortage?
>
> A. What cars should people buy?
> B. Are car makers building fewer new models?

C. Are people using less gasoline?
D. How large should cars be?"

The above item purports to test Functioning Within Social Institutions (student knowledge of the relationship between social institutions and individuals) and Solving Problems (i.e. the process by which a problem was solved).

Validity and Reliability

Internal consistency reliability for the Composite Examination based on comparisons of equivalent forms has been determined at .86 and .89 for the total scores for two different samples of students. Sub-score reliabilities ranged from .58 to .80. Reliability estimates for individual scores on the Objective Test were .84 for the Total score and only .63 to .68 for the six sub-tests. Reliability estimates for group mean scores on the Objective Test were .92 for the total score and .86 to .92 for the six sub-tests (Steele, 1988). Thus reliability estimates for group total and sub-test scores on both types of COMP examinations are sufficiently high; individual scores on the six sub-tests are, however, not reliable.

A test of the degree to which students will respond to the Composite Examination in the same way on two different occasions within a reasonably short period of time (test-retest reliability) also produced acceptable reliabilities, as did inter-rater reliability, which is important only for the free-response items of the Composite Examination. Modest differences in the level of ratings, however, led ACT to refine items and faculty-rater procedures.

In sum, the reliability estimates for individual scores on the sub-tests of the Composite Examination and the Objective Test are not high enough for use in student certification or individual placement. For program evaluation, which is the primary use of COMP and for which group mean scores are used, the reliability estimates for both the total score and the sub-scores are sufficiently high.

Validity

Do the sub-scales measure what they were intended to measure and are they reasonably independent of each other? As mentioned earlier, several of the items were difficult to classify, especially for the Solving Problems and Clarifying Values sub-scales. These two sub-scores intercorrelated .58 in one sample and .57 in a second sample (Forrest and Steele, 1982). Other scales intercorrelated in the .51 to .59 range as well. These correlations for the Objective Test were exceeded on the Composite Examination. In fact, the Solving Problems/Clarifying Values intercorrelation was as high as .77 for one sample. Correlations of this magnitude indicate that the scales are measuring overlapping areas, especially on the Composite Examination.

The COMP developers made two major assumptions about the uses of the COMP instruments:

> (1) the outcomes assessed by COMP are relevant to effective functioning in a variety of adult roles in civic and work settings; and
> (2) these outcomes are amenable to instruction and are being developed in college settings (Forrest and Steele, 1982, p. 33).

To evaluate the relevance of COMP measures to effective functioning as an adult, the COMP staff used employee supervisor ratings as one criterion of effectiveness. Multiple correlations with COMP sub-tests were in the .40 to .60 range for several samples of adults who had been out of college less than a decade (Forrest and Steele, 1982). These indicate a fairly high relationship between a combination of the COMP sub-tests and ratings by supervisors. The multiple R's were much lower for samples of adults who had been out of college longer than 10 years, suggesting that other job-related or personal factors may affect older adult functioning. Although several of the studies suggest a sizeable amount of variance in job performance explained by COMP scores, traditional academic achievement or aptitude measures may also have had significant correlations with effective performance. The studies would have been strengthened by statistically controlling for general academic ability (e.g. ACT or SAT scores).

The more important validity issue is the relevance of the COMP measures to student achievement in general education programs in colleges and universities. The COMP staff conducted a series of studies with the pilot group of colleges to examine this issue (Forrest and Steele, 1982). Among the findings:

1. Sophomores and seniors score higher than freshmen. The greatest score "gains" (defined below) occur by the end of the sophomore year. But because these were not matched samples, the higher upperclass student scores are likely due in part to the attrition of lower ability students from the sample. In another study of two groups of institutions that required different amounts of general education coursework, the average score "gain" was twice as large for the institutions that allotted more of their B.A. program to general education.

2. Seniors did better on the sub-test most appropriate for their field of study (e.g. social science majors on Functioning Within Social Institutions). A more

significant validity finding was reported at Bemidji State University: In two of the three sub-related areas of COMP, seniors who were non-majors did better if they took even a few more courses in the area. The relationship, however, was not strong and may be related to initial student interest and background.

In sum, the COMP measures appear to be reasonably related to student performance in many general education courses. But how appropriate are they for assessing programs in a selective college? A study at the University of Minnesota concluded that the COMP exam was too easy to differentiate among students of high ability (Schomberg, et al., 1981). If this is the case, institutions with large numbers of high ability students would not be able to show growth for seniors due to a ceiling effect. A recent analysis of colleges with varying levels of student ability which had retested students two or four years later, however, found gains in every level of entering achievement (Steele, 1988). More analyses of the performance of high ability students on the COMP test are needed.

COMP Scores as Measures of Estimated Gain

Institutions requiring ACT Assessment scores for admissions may use them to estimate the mean total on the Composite Examination for entering freshmen. Studies of the relationship between the ACT Assessment and COMP scores for large samples of students were the basis for the estimations. After administering the COMP to sophomores or seniors, institutions would then be able to estimate gain on the COMP. ACT publishes "concordance" or expectancy tables for institutions to estimate the earlier scores (total score equivalents only).

Studies at the Learning Research Center at the University of Tennessee, Knoxville (UT-K), found that estimates of student score gain on the COMP can be greatly in error if a large portion of students do not have ACT Assessment scores, because those who do have scores are a nonrepresentative sample. Use of the COMP concordance table also resulted in a significant over-estimate of the freshman COMP total score, and therefore an under-estimate of gain at the University (Lambert, 1985; Banta, et al., 1987). As with change scores generally, the estimated gain score is less reliable than either of the tests used in computing gain (the UT-K researchers reported reliability coefficients of .44 and .54). A final problem cited by the UT-K researchers was the validity of estimated gain as a measure of an institution's general education program. Estimated gain scores were not logically related to program characteristics or student activities that reflect good educational practice. For example, estimated score gain was higher for students who took no honors sections, who did not participate in the University's freshman orientation program, and who took little or no natural science or math. Such findings suggest no positive actions that an institution could take to improve its general education program.

Actual gain rather than estimated gain scores would be a preferable approach to studying program quality. That is, by giving one form of the test to incoming freshmen and another form to the same (remaining) students at the end of their sophomore or senior year, there is no need to estimate the earlier set of scores. This longitudinal design also provides for sub-test analysis whereas the use of concordance tables only provides total score estimates. Sub-test analysis can reveal areas of strengths and weakness and may therefore be more useful for program assessment. Nevertheless, the longitudinal design will likely be used sparingly because of the time lapse required, the need to motivate students to do their best at both testing times, and the fact that gain scores still have imperfect reliability.

The Academic Profile

The Educational Testing Service (ETS) is currently pilot-testing the Academic Profile as a new assessment service for general education. The Academic Profile supercedes ETS's Undergraduate Assessment Program (UAP), no longer published. The UAP was designed to "provide a measure of students' knowledge and grasp of basic concepts in the broad areas of the liberal arts," and included Area Tests in social science, humanities, and natural science.

The same three discipline areas--social science, humanities, and natural science--form the context of the questions in the Academic Profile. The matrix is completed with the assessment of four academic skills--college-level reading, college-level writing, critical thinking, and using mathematical data (see Figure 1). These four skills were drawn from the 1985 Association of American Colleges report, Integrity in the College Curriculum.

As indicated in Figure 1, the Academic Profile and the COMP matrices appear similar except for the Clarifying Values scale in COMP and the Using Mathematical Data in the Academic Profile. Also, the COMP emphasizes the use of social science, natural science, and humanities material in adult roles (e.g. Using Science and Technology), while the Academic Profile is more traditional and curriculum-related.

The stimulus materials in the Academic Profile include a passage, poem, graph, or table followed by a question or questions based on the stimulus. For example, a brief poem is followed by questions testing reading ("Which statement best expresses the idea of the poem?"), writing ("Which of four words could be inserted in a given line without changing the meaning?"), and critical thinking (comparisons between two concepts in the poem). Another set of items is based on a four-paragraph modern history passage that purportedly tests reading and critical thinking. As with the previous poem-based example, the critical thinking questions seem to be heavily dependent on reading comprehension skills, so the two sub-scores may prove to be highly inter-correlated. Reliability and validity information has not been published on The Academic Profile, but is expected to be based on the pilot testing. Although no data are yet available, the questions appear to be somewhat more difficult than those of the COMP examinations.

An optional 45-minute essay in which students respond to a sample question in one of the three discipline areas of their choice is also available. The essays are to be scored by the institution, using manuals provided by ETS, with instructions for both holistic and analytic scoring. To produce reliable results for individual assessment, institutions will likely have to train faculty members.

A short form (one-hour testing time) consisting of 48 items provides group scores only. A long form (three hours) provides both group and individual scores. The short form is "spiraled" for distribution, meaning that there are three forms of the test, and in a group administration, a random third of the students takes each form of the test. Group scores are reported for the 144 (3x48) item test, resulting in more information for curriculum analysis.

Institutions may add up to 50 locally written questions to be scored by ETS. This option allows institutions to test at least some of their specific curriculum content. If carefully constructed and pre-tested, the locally produced items could be a useful supplement to the eight scores (seven scales plus a total) provided by the Academic Profile.

As with the COMP program, ETS will provide expectancy tables using the Scholastic Aptitude Test (SAT) or ACT Assessment freshman scores to estimate Academic Profile freshman level performance. That is, based on the likely positive relationship between SAT and Academic Profile scores (or the ACT Assessment and Academic Profile scores), it is possible to estimate how entering students would score on the Academic Profile. Sophomore or senior Academic Profile scores can then be compared to the estimated freshman scores. It is likely that the problems found with gain scores on the COMP will also occur with the Academic Profile.

The College-Level Examination Program (CLEP)

The CLEP tests are administered for the College Board by ETS. Five CLEP tests were designed as certification exams to help students meet general education distribution requirements.

The five (English Composition, Humanities, Mathematics, Social Sciences and History, and Natural Sciences) are each 90-minute long multiple-choice tests based on course work during the first two years of college. At the current rate of $12.00 to $14.00 per test and at 90 minutes per test, for a college to use all five tests to assess their general education program offerings would be both costly and time consuming. Given these limitations, the College Board recently has indicated a willingness to work with institutions to provide shorter versions of the examinations for program assessment.

The major use of the CLEP General Examinations has been to provide college credit to traditional and non-traditional students who mastered course content outside of college. Norms based on college students who have completed general education course work are provided, but each institution sets its own cutting scores and determines what amount of credit should be given the tests. Guidelines have been suggested by the ACE Commission on Educational Credit and Credentials (Whitney and Malizio, 1987). Reliability coefficients for the tests are all above .90 and although scores are reported on the familiar 200-800 scales, they should not be equated with other College Board tests. The explanatory material provided with each test lists the content covered, such as mechanics of writing in English composition; and literature, art and music in the humanities test. Colleges may not find that the General Examinations adequately cover all of the desired content areas (see Adelman's essay on difficulty levels in this volume), and there is no doubt that the General Examinations, compared to the CLEP Subject Examinations, call for less depth of attainment in each discipline. Since the Subject Examinations are also used to grant credit to students, institutions could choose to have students satisfy distribution requirements in general education with selected Subject Examinations.

Locally-Developed Measures

As stated at the outset, faculty and staff involvement in the assessment process is necessary if the results are to have maximum potential for changes in the curriculum and teaching. If commercial tests are used to assess student learning, faculty involvement in the selection and review of the tests, as well as in the interpretation and application of the results, is essential.

But faculty can also design their own measures and score the responses. The major advantage in this approach is that it provides faculty members with the opportunity to clarify what they want students to learn in general education and to formulate questions and desired answers to test those expectations. A disadvantage is that faculty members may not have the time, commitment, or expertise for such an undertaking. But at least one major multi-institutional experience, described next, belies that conventional wisdom.

Academic Competences in General Education

The goal of this FIPSE-funded project was to involve the faculties at 15 California colleges in developing free response questions to assess competence in four selected areas of general education: communication skill, analytic thinking, synthesizing ability, and social/cultural awareness (Warren, 1978). These four areas were selected and defined by a committee of faculty from the project colleges after considering a number of descriptive statements of college student competencies and course objectives derived from an earlier study (Warren, 1976). Communication involved only written expression, analytical thinking included cognitive operations such as discrimination of fact from conjecture and the relevant from the irrelevant, and awareness included the importance of values in human affairs. To measure the four areas, questions were written in four content fields: natural science and mathematics, social sciences, humanities, and history and political science.

Using the above matrix specifications, faculty members at the project colleges wrote free-response questions that students could answer in less than ten minutes each. These were administered to a large sample of students at the colleges. Categories for scoring were developed inductively on the basis of actual student responses. Faculty members sorted responses to a question into four to nine categories that evidenced different types of performance on the competence being assessed. Categories were then described, revised and cross-validated by other faculty scorers in order to determine agreement among the categories.

The reliability of group scores, as reflected by the extent of agreement among scorers, was acceptable for most of the questions. That is, if the questions are used to assess the performance level of groups of students for program or curriculum evaluation, then acceptable levels of reliability were achieved. The reliability level for individual scores, however, was too low for accurate assessment of individual student learning.

Involving themselves in the scoring of questions helped faculty members clarify their own thinking about the kinds of learning general education should address; faculty members also became better versed about the kinds of deficiencies that were most prevalent in their students. When the questions were later tried at institutions outside the project, however, faculty members reported that many of the questions were too difficult for their students and too complex to score. Both of these criticisms could, of course, be addressed in future efforts by college faculties. Examples of the content of questions, the area assessed, and the major response categories are provided below. (Appendix B of the project report provides further examples).

Example A. To assess communication (with humanities content), students are given a one paragraph synopsis of The Cask of Amontillado written in eleven short, choppy sentences. They are asked to rewrite the passage in three or four smooth sentences that retain the important content.

Four levels of performance criteria were established for this item:

1. The paragraph is concise, accurate, graceful, and essentially complete.
2. The paragraph meets three of the four requirements above.
3. The paragraph meets two of four requirements.
4. The paragraph meets fewer than two requirements.

Example B. To assess awareness (with history and political science content), students are asked to draw some inferences about the society in which the following verse was written:

"Though I am but poor and mean,
I will move the rich to love me,
If I'm modest, neat and clean,
And submit when they reprove me."

Six levels of performance criteria in three general categories were used for this item:

1. The response makes an explicit statement about one or more values implied in the poem--propriety, submissiveness, material wealth--going beyond the specific words of the poem, and at least implies the existence of a strong class structure.
2. The response points out the class distinctions but makes no reference to values.
3. The response focuses on the writer's individual position without mention of the society in which he or she writes.

The forty-seven questions written for the project yielded scores only in the four competencies: communication, analytic thinking, synthesizing ability, and awareness. There was no intention of building reliable sub-scales in the four academic areas but such an effort could be part of a faculty project to build on this model. Warren's analyses did not include information on the correlations among the four competencies or other validity data. The extent to which the scales were independent measures and did in fact test analytic thinking, for example, was not ascertained. The test did, however, appear to have face validity as well as group score reliability.

Other institutions, such as Kean College of New Jersey, have recently undertaken major local efforts to assess their undergraduate programs. At Kean the emphasis is on faculty developed, criterion-referenced measures that are tied to course and curriculum objectives at the college (Presidential Task Force, 1986). Another strategy followed by some colleges is to have the course syllabi and the final examinations in key general education courses assessed by outside readers who respond to a carefully structured questionnaire.

Conclusion

As Warren once observed, "the problem with general education is its generality." There are as many conceptions of this critical portion of the college curriculum as there are faculty councils and curriculum committees to bring them forth. Nonetheless, there are enough common underlying dimensions of the major types of general education programs for valid assessments to be designed and used with reliable results. The challenge to faculty is to identify those dimensions, examine existing instruments to determine their appropriateness to both curriculum and student abilities, and select an assessment scheme that meets the institution's need for helpful information within the constraints of available resources (principally faculty time and direct testing costs). For purposes of program evaluation, it may very well be that faculty select more than one instrument or method for the assessment of general education, using a matrix sample of students, thus providing a richer description of program effects and mitigating the problem of "generality."

Indeed, when States initiate assessment programs for public institutions, they usually advocate the use of multiple indicators of student learning in one clause and faculty responsibility for designing and carrying out assessment in another (Education Commission of the States, 1987). Given the reviews of the major commercial instruments and models of local measures cited in this essay, a faculty can begin to meet both suggestions within the context of general education.

References

Advisory Committee, College Outcomes Evaluation Program, Report to the New Jersey Board of Higher Education, COEP Program, Trenton, N.J.: New Jersey Department of Higher Education, 1987

Alverno College Faculty. Liberal Learning at Alverno College Milwaukee, Wisc.: Alverno Productions, 1976.

Banta, T.W., and Fisher, H.S. "Measuring How Much Students Learned Entails Much More Than Testing Them," The Chronicle of Higher Education, March 4, 1987, pp. 44-45.

Banta, T.W., Lambert, E.W., Pike, G.R., Schmidhammer, J.L. and Schneider, J.A. "Estimated Student Score Gain on the ACT COMP Exam: Valid Tool for Institutional Assessment?" Paper presented at the Annual Meeting of the American Educational Research Association, Washington, D.C. April, 1987.

Education Commission of the States, Assessment and Outcomes Measurement: A View from the States. Denver: Author, 1987.

Forrest, A., and Steele, J.M. Defining and Measuring General Education Knowledge and Skills. Iowa City, Iowa: The American College Testing Program, 1982.

Grant, G. and Associates. On Competence: A Critical Analysis of Competence-Based Reforms in Higher Education, San Francisco: Jossey-Bass, 1979.

Klein, T.D. and Gaff, J.G. Reforming General Education: a Survey. Washington, D.C.: Association of American Colleges, 1982.

Lambert, W. Problems with Estimated Gain. Knoxville: Learning Research Center, Univ. of Tennessee, 1985.

Levine, A. Handbook on the Undergraduate Curriculum. San Francisco: Jossey-Bass, 1978.

Marton, F. "Skill as an Aspect of Knowledge." Journal of Higher Education, Vol. 50, no.5, 1979, pp 602-614.

Mentkowski, M., and Doherty, A. Careering After College: Establishing the Validity of Abilities Learned in College for Later Careering and Professional Performance. Final Report to the National Institute of Education. Milwaukee, Wisc.: Alverno Productions, 1983.

Mentkowski, M. and Loacker, G. "Assessing and Validating The Outcomes of College," in P. Ewell (ed.) Assessing Educational Outcomes. New Directions in Institutional Research, No. 47. San Francisco: Jossey-Bass, 1985, pp. 47-64.

Presidential Task Force, Kean College, "A Proposal for Program Assessment at Kean College of New Jersey: Final Report." Union, N.J.: Kean College of New Jersey, 1986.

Schomberg, S.F., Hendel, D.D., and Bassett, C.L. "Using the College Outcome Measurement Project to Measure College Outcomes." Paper presented at the Annual Forum of the

Association for Institutional Research, Minneapolis, 1981. ERIC Document No. ED 205121.

Steele, J.M. "Using Measures of Student Outcomes and Growth to Improve College Programs," Paper presented at the Association for Institutional Research, May, 1988.

Warren, J. "College Grading Systems." In Goodman, S.E. (ed.), Handbook on Contemporary Education. New York: R.R. Bowker, 1976, pp. 172-175.

Warren, J. The Measurement of Academic Competence, Final Report, Grant #G007603526, Fund for the Improvement of Post-Secondary Education, Dec. 1978.

Whitney, D.S. and Malizio, A.G. Guide to Educational Credit by Examination. New York: Macmillan Publishing Co. and the American Council on Education, 1987.

Assessment Through the Major

by Mark I. Appelbaum

For many years the informal assessment of the quality of programs in higher education has been concentrated at the level of the major, department, or school. For instance, the subjectively ranked quality of graduate education in the non-professional schools (e.g., the Rouse-Andersen ratings, 1970) is always reported by department or major. The quality of graduate professional schools has frequently been judged (particularly by groups such as boards of governors, State legislatures, and senior administrative officials), at least in part, by passing rates on disciplinary credentialling examinations such as State Bar Examinations and Medical Board Examinations. Similarly, the quality of undergraduate professional programs such as those in nursing and education are assessed, in part, by passing rates on certifications such as state licensing tests in nursing and the National Teachers Examination.

The use of the department or major as the unit of assessment has several distinct advantages over an institution-wide assessment scheme which attempts to assess the quality of undergraduate education in toto. These advantages include:

(1) the size of the assessment project itself, with a small number of well conceived and designed assessments phased in each year;
(2) the greater possibility of a proper fit between the form of the assessment and specific features of the instructional program; and
(3) the close connection between the assessment unit and the instructional unit.

Let us briefly examine each of these advantages.

Two concerns dominate the planning of assessment programs in higher education: the quality of the assessment itself and its acceptability to both the faculty of the institution and the eventual consumers of the information it yields. By basing the assessment at the level of the department (or major) it is possible to phase in the assessment process in stages. The selection of the first participating programs should be based on a combination of factors including the perceived quality of the department or major, the potential acceptance of the assessments by the faculty involved, the experience of the field with such assessments, and the technical competence of involved faculty to launch such an assessment. By selecting the initial participants so that the first experiences with this type of assessment are as positive as possible, the overall acceptance of such assessments by the faculty in general may be enhanced.

The second, and perhaps most important, advantage of basing the assessment program at the major or department level is that of optimizing the fit between the content and form of the assessment and the particular goals and sensibilities of the instructional unit. There is no reason to assume that every department has the same instructional objectives. Some may emphasize the knowledge base of the field, while others may emphasize the role of theory; some may emphasize performance and production, and still others may emphasize the role of empirical research. Some departments may build their curriculum around a narrow set of required courses and prerequisites while others may offer a curriculum with few requirements and many electives. Some departments may structure their programs around formal lecture/discussion courses while others may provide a substantial amount of field experience, research experience, individualized reading courses, or other alternatives to formal classroom instruction. Such differences clearly exist between disciplines within an institution as well as within disciplines across institutions.

For example, psychology departments vary widely in the role of field and clinical experiences within the curriculum, in the number and types of laboratory experiences available, as well as in their orientation towards clinical, cognitive, and biological issues. And if the course-taking patterns of majors are influenced by particular faculty (e.g. a leading authority on Chinese history), then a department's expectations for student learning will take on a different configuration than the "core" of most history curricula. By basing the assessment program within the major or department, such differences in approach and expectations can more easily be recognized and the assessment designed accordingly.

Basing the assessment at the departmental level also has the advantage of linking the assessment unit and the instructional unit. This linkage is clearly advantageous should the evidence produced by the assessment indicate that the instructional program is operating at less than an optimal level. Since the content and form of the assessment would have been established by the instructional unit itself, there could be little convincing argument that the problem is with the assessment rather than with the instructional program. Further, the involvement of faculty in the design and implementation of the assessment per se should provide some motivation towards rethinking and reevaluating instructional objectives.

Issues in Designing the Assessment

Should the decision be made to base the assessment within the individual department or major, a number of issues must be considered in the early phase of design. These include the purpose of the assessment, content and coverage, form and format, sampling and motivation of respondents, timing of the assessment, and the technical characteristics of the assessment instrument and methods, including a consideration of the "norms" to be applied. While each of these issues will be explored separately, one must recognize that they are not independent and that decisions made in one forum will have implications for the others.

The Assessment Team and Its Charge

It is essential that each of these issues be considered by that group of individuals charged with the responsibility of designing, implementing, and reporting the results of the assessment. This group of individuals must be a carefully selected team, collectively knowledgeable about the instructional goals of the unit, current directions in the field, and technical knowledge of the principles of evaluation. While most members of this group should be faculty in the department being assessed, it would be wise to include at least one senior faculty member from outside the department but in a related area, one individual experienced with the administrative sensitivities of the institution, one person experienced in the assessment field if no such individual is available from the target department, and at least one advanced undergraduate in that discipline. The advantage of creating a fairly broadly based assessment team is that while the disciplinary concerns of the area being assessed are represented by those faculty in the department, the presence of individuals outside of the unit (including a student) prevent the assessment from being excessively parochial or self-serving.

It is also essential that this panel be clearly charged by those with the overall administrative responsibility for the assessment as to its purpose, scope, financial limitations, and the expectations the administration may have for the assessment process. Included in the charge must be a specification of how the results of the assessment are to be reported and to whom.

It may well be the case that several forms of an assessment report will be necessary--one for the student participant in which her or his individual performance is reported (with relative strengths and weaknesses noted); one for the faculty in which detailed analyses of the strengths and weaknesses of the curriculum are noted along with suggestions for changes; and one for the senior-level administrator which accurately and succinctly summarizes the findings of the assessment and recommendations for improvements in the curriculum. In this latter report it is important that the assessment procedure itself be described in some detail so that the process can be evaluated by those receiving the results. Indeed, the senior-level administrator responsible for the several assessments should ask for interim reports detailing the assessment plan prior to its implementation to ensure prior agreement on its adequacy.

All too often assessment projects are seen by faculty as time-wasting exercises with no real consequences other than to fill filing cabinets in the administration building. Thus, in addition to charging the assessment team with regard to the purpose of the assessment, the appropriate academic administrator should give clear indications of how the assessment results are to be used.

Purposes of the Assessment

The first issue that must be clearly specified and understood by all parties involved in the design and execution of the assessment is the purpose for which it is being conducted. The first of these purposes is to audit the content of the curriculum and to assure that all aspects of the discipline that ought to be offered are offered, and that a suitable proportion of students who graduate in that "discipline" have been exposed to that content. A second purpose of the assessment may be to provide some measure of the efficacy of the undergraduate instructional unit (be it a single department or an interdepartmental "committee"). An evaluation at this level would necessarily involve the assessment of not only what is taught, but also how well it is taught, retained, and internalized by the "typical" or "representative" student in that discipline. A third purpose for which an assessment might be designed is the certification of the individual student, i.e., to measure the degree to which each student in the field has mastered the objectives of the curriculum and to provide some qualitative or quantitative index of the degree and level of that mastery. Such an index might be employed to determine minimal competence for graduation (or honors).

It is essential for those ultimately responsible for the conduct of the assessment to understand that these purposes are quite different (although related); that for each purpose a different form of assessment with rather different considerations (both technical and pragmatic) would be necessary; and that the results of one form of assessment cannot be utilized directly for one of the other purposes.

Content and Coverage of the Assessment

If the content of the assessment does not correspond to the agreed upon goals set by the faculty, the results of the assessment surely will be dismissed as irrelevant to the instructional program. In the following discussion it is assumed that the assessment will focus on the knowledge obtained as a result of instruction in the discipline and that the mechanism will be some form of "testing," be it multiple choice, essay, or some other format. Should it be decided that some alternative format, such as a "senior essay," experimental project, position paper, or examination by an external visiting committee be adopted, then other considerations of content and coverage would need to be applied. But assuming that the test mode would be the most prevalent format, the following issues should be considered carefully.

First, the scope of coverage. It might be decided that the assessment should be based upon the entire field as it is conceived by the mainstream of the profession. This scope of coverage is perhaps the most commonly used and represents the type of knowledge testing employed throughout the educational community when it elects to use commercially developed tests at any level. For example, each of the Graduate Record Examination Subject Area tests consists of a set of items determined by a panel of disciplinary experts to represent a field broadly conceived, but without regard to the course offerings in any particular department or institution. The utility of this approach has been discussed by McCallum (1983) in the context of the evaluation of a psychology program, but its limitations have been sufficiently recognized (see, e.g., Burton, 1982) so that ETS has now developed an alternative: the Major Field Assessment Tests (MFATs).[1]

A limited number of professional associations (e.g., the American Chemical Society) have developed examinations to assess the knowledge base of the field as defined by the professional association. This latter form of assessment device has often been developed in conjunction with the association defining a model curriculum in the field or responding to traditional divisions of knowledge within a field (e.g., organic chemistry, physical chemistry, etc.). The advantage of such an approach is that it assesses the instructional program in the context of the field and tends to ensure that the local curriculum has not become either dated or extremely idiosyncratic. Further, this approach allows the possibility of the use of norms (see below) which can be based on more extensive samples than any locally developed test. The major disadvantage of such an approach is that it cannot be sensitive to the locally established goals of the instructional program and can easily ignore major thrusts of the curriculum being assessed. Also, this approach tends to under-evaluate small programs which simply cannot provide coverage of all aspects of the field.

A second approach to the scope of coverage issue is to use the full array of course offerings in the department being assessed to define the content domain of the assessment and to then sample items from that domain (see below) thereby producing a "comprehensive" examination as defined by the courses in the department. Banta and Schneider (1986) discuss the development of such comprehensive examinations as part of an institution-wide effort at the University of Tennessee.

Finally, one might decide to take a more "minimal competency" approach to assessment at the departmental level by defining a smaller set of courses (perhaps the core courses or areas required of all majors) as the domain of coverage. Under this approach one might expect to find a narrower but deeper form of testing. In a limited number of cases, course-based tests such as those developed by the American Chemical Society (see above) could be employed as long as the tests reasonably correspond to the content of the selected courses. Should one decide to take the "minimal competency" approach, it is critical that the defined minimum be rich enough so as not to trivialize the assessment program. Minimal breadth does not mean minimal depth.

The problem of the scope of coverage becomes more complicated in the case of interdepartmental B.A. programs such as Area Studies, Criminal Justice, Comparative Literature, and so forth. In these cases, special attention needs to be directed to the assessment of the curriculum offered by a set of faculty members specifically designated as the core faculty, as well as the assessment of those areas of coverage provided by the other faculties allied with the core curriculum.

In those disciplines in which "outside of classroom" activities such as field placements and internships are considered to be essential features of the curriculum, special attention should be given to the manner in which the results of these activities are to be built into the assessment. An example of the assessment of one such activity can be found in Morris and Haas (1984).

Determination of Content

A starting point for determining of the content of the assessment instrument would be an explicit statement of objectives by the instructional program faculty. This process is difficult and time consuming, but of great importance and benefit. (It well may be the first time that some faculties have collectively discussed the objectives of their total educational program.) From the resulting grid of curricular objectives, a subset must be selected depending upon the decision that has been made with regard to scope. From this point, the actual development of items can be undertaken by (a) writing items *de novo* that match the curricular objectives selected, (b) selecting items from existing examinations or other item pools available within the unit, (c) selecting a commercially available test that has a high content overlap with the curricular objectives selected (say an 80 percent overlap), or (d) creating a hybrid instrument. This last approach is one in which a set of items from tests with known properties (such as norms) is imbedded within a batch of locally developed items. This use of reference items, as they are technically known, allows a partial comparison of the results of a local assessment with those of other assessments utilizing the same reference items. Once problems of copyright are solved, the use of a hybrid instrument mitigates a number of the technical problems cited in this essay.

No matter which of these approaches is preferred, the selection of particular items (or tests) should be informed by a content representativeness study. The basic idea underlying this type of study is the formation of a grid which has as its columns the specified objectives of the instructional program and as its rows the items under consideration, be they an item pool or the items on an established test. The task in such a study is to then determine the match between each item and the stated objectives. This grid also may be used to discard items which do not measure any of the department's curricular objectives and to generate new items to measure objectives for which no items yet exist. The same approach can be taken with essay and other performance assessments, though the task is more difficult. Generally, the more restricted the possibilities of student response (as is the case with test items requiring choice or completion) the greater the precision of a content representativeness study.

Examples of this approach can be seen in a recent series of studies sponsored by the Office of Research of the U.S. Department of Education. These studies evaluated some of the leading content area examinations in selected fields against surveys of faculty and professional consensus on the objectives of undergraduate learning in the major, and offer a useful starting point for considering commercially available instruments for assessment.[2] In addition, some test publishers have sponsored content representativeness studies of their tests which may also be of use. For instance, Oltman (1982) and DeVore and McPeek (1985) offer content assessments of several of the GRE Area Examinations (specifically biology, literature, political science, chemistry, education, and computer science). These content assessments, however, are stated in rather general terms (e.g. the categories of content representativeness for the Computer Science Area Examination are Software, Systems and Methodology, Computer Organization and Logic, Theory, Computational Mathematics, and Special Topics) and are not substitutes for a detailed examination of the individual instruments by a local faculty. On the other hand, the methodology described in both of the papers is valuable and should be considered by the assessment team.

Form and Format of the Assessment Instrument

Having decided on content, one next needs to consider the administrative form and test format of the examination itself. There are a number of alternatives. Concerning administrative form, one could elect a single sitting examination (traditionally a three-to-four-hour exam--a period thought to be near the limit of good student performance) in which all of the agreed upon content is examined. Alternatively, one could assess subsets of the agreed upon content in shorter sessions spread over a longer period of time. A third alternative is one in which a single session is scheduled, but in which a multiphasic battery of tests is employed. Under this plan a potentially different set of sub-tests is administered to each student based upon the specific courses the examinee has taken.

A variant of this third scheme might prove useful if it is determined that the domain of coverage is too great to allow a single test to be administered to all examinees. Under this variant, random subsets of the total item pool are gathered together to form several weakly-parallel forms of the test, and each is administered to a different subset of examinees (see Millman's discussion of matrix sampling, above). While this system of testing is generally not accepted for purposes of individual assessment (i.e. certification), it has proven useful in a number of program assessment applications, most notably in the recent rounds of the National Assessment of Educational Progress (Beaton, 1987). This system has the additional advantage of maintaining test security in multiple administrations of the assessment.

A second and more serious decision concerns test format. There are many formats, each with its own distinctive advantages and disadvantages. The decision on test format will undoubtedly have many consequences ranging from the number of students who can be tested to the cost of the assessment process. The most common format for assessment of this type is the multiple choice format. The primary advantages of the multiple choice format are its familiarity, the precision and ease of scoring either by hand or machine, and the ease of establishing the psychometric properties of such tests. The net result of these advantages is that the assessment can be quickly, accurately, and inexpensively administered once the test has been constructed, and consequently a large number of students can be included in the assessment. One further advantage of the multiple choice format is that most (but certainly not all) commercially available assessment devices, such as the public release forms of the GRE subject area tests, are written in this format. (The public release forms of the GRE change fairly often and, if employed as part of an assessment program, it is desireable to use the most recently released form if for no other reason than to minimize the chances that students will have seen the items as part of their preparation for taking the GRE.)

There are a number of serious disadvantages to the multiple choice format. The most serious of these disadvantages deals with the level (in terms of cognitive demands) at which multiple choice items tend to operate. Bloom's Taxonomy of Educational Objectives lists knowledge (recall and recognition), comprehension, application, analysis, synthesis, and evaluation as the goals of the educational process in increasing order of cognitive demand. Analyses of multiple choice items have repeatedly shown that this form of test item rarely, if ever, operates beyond the level of simple recall and recognition. The problem with multiple choice items seems to be inherent in the item type itself, for when attempts have been made to rewrite items to require a higher level of cognitive functioning (in Bloom's sense) experienced item writers were unable to produce substantial numbers of items which operated beyond the recall/recognition level (Levine, McGuire, and Natress, 1970). As Lyle Jones has noted, "There is evidence that the form of the multiple choice item is intractable with respect to measuring higher cognitive skills" (Jones, 1987).

Clearly, the degree to which multiple choice format examinations pull for "recall and recognition" level skills is somewhat a function of the area being tested. In some areas (e.g., chemistry, physics, and the various specialties in engineering) it has been well demonstrated that multiple choice items that demand higher level cognitive skills, such as analysis and synthesis, can be constructed. (This is not to imply, however, that commercially available tests in these areas tap these higher order skills.)[3]

The concern that commercially available multiple choice format examinations do not tap a proper level of cognitive functioning is clearly expressed by Peterson and Hayward (1987) in their "A Review of Measures of Summative Learning in Undergraduate Biology." After reviewing fourteen instruments purported to examine general biology, they remark:

> We are alarmed and perplexed that there appear to be no standardized tests available in the U.S. that attempt to determine the degree to which graduates are able to function as scientists, or more specifically as biologists. That is, existing tests do not measure whether an individual can identify a problem, ask a research question, develop hypotheses, test them, or draw a conclusion and report the results.... The evaluation of learning pertaining to the process of scientific inquiry may be severely limited as long as we rely on multiple-choice tests as the principal method of measuring achievement.

The second disadvantage of the multiple choice item for the type of assessment envisioned is actually an interaction of the problem of the "cognitive demand" level of multiple choice items with a fear that testing drives instruction and the curriculum. This concern is perhaps best summarized in the following brief statement from the Committee on Research in Mathematics, Science, and Technology of the National Research Council (1987, p.20):

> Most present classroom methods of testing what students know emphasize the recall of facts--as does most teaching. If tests are not to trivialize instruction further, new approaches to assessing student achievement must be developed that aim at conceptual understanding, the ability to reason and think with scientific or mathematical subject matter, and competence in the key processes that characterize science or mathematics.

Should the assessment instrument consist mainly of items that demand only recall and recognition and should the assessment have any pragmatic impact, the net result might be to dilute--rather than to improve--instruction. A somewhat more optimistic view of the potential for multiple choice items and the conditions under which this optimism might be realized can, however, be found in Frederiksen (1984), Frederiksen and Ward (1978), and Ward (1986). Their work indicates that items (both multiple choice and brief answer formats) can be developed that tap scientific creativity and problem-solving abilities. The dominant format of experimental items developed by Frederiksen and Ward involves the presentation of a situation with tabular data such as the relationship between birth weights and I.Q. scores or the capacity, patient turnover and death rates in a hospital in 18th century London, and asks the student first to generate a series of hypotheses to account for the major relationships among the data and then to indicate the most likely hypothesis. This format combines features of a free-response question with those of a restricted-choice question.

Other formats traditionally used in classroom assessment are obviously available. These include open-ended and free-response questions which require the examinee to generate a correct answer as opposed to simply recognizing one; essay questions which further require the examinee to analyze, synthesize, and organize a body of knowledge as opposed to a single fact or issue; or even a production task which might be designed to assess the student's ability to apply a body of knowledge to a well-structured problem or (perhaps even more importantly) to an ill-structured problem.

While each of these approaches allows the assessment of higher level cognitive domains, they are not without some limitations. These limitations include the substantially higher cost of administering and evaluating the assessment, the greater subjectivity attendant upon their scoring and interpretation, and the lack of familiarity that most academics have with the technical demands of such forms of assessment (despite the fact that they frequently use them in the classroom). Nonetheless, if the goal of the instructional program is to achieve excellence in the use of these higher order skills within the context of the discipline, the designers of the assessment must carefully consider the format of the instrument and its implications in light of that goal.

Other Forms of Assessment

The system of examination commonly employed in the British university and polytechnic system merits some attention at this point. This system, discussed at some length by Lawton (1986), Sawyer (1976), Tannenbaum (1986), and Williams (1979) among others, combines an extensive written senior comprehensive examination (a series of extended essays) with the use of an examination board that includes at least one external examiner. While the basic purpose of the system is to award degree levels, it also achieves certain of the goals of assessment envisioned in this volume including assessment of students on higher order cognitive levels. Additionally, the presence of the external examiner allows for at least a partial assessment of the overall program of instruction across institutions. As noted by the Committee of Vice Chancellors and Provosts (1986):

> The purposes of the external examiner system are to ensure, first and most important, that degrees awarded in similar subjects are comparable in standard in different universities in the United Kingdom, though their content does of course vary; and, secondly, that the assessment system is fair and is fairly operated in the classification of students.

Experiences of American universities and colleges with the external examiner system are detailed in Fong (1987). There is a fundamental difference between the two systems. The British external examiner is essentially an auditor and mediator, functioning at the second level of review. In the few cases of external examining in the U.S. (e.g. in Swarthmore's honors program), the external examiner functions at the first level, writing and grading senior examinations, and possessing the sole authority to recommend honors. For this system to work in relation to local curricula and local faculty expectations for student learning, there must be a high degree of explicit prior consensus between the external and internal faculty on performance objectives.

The designers of large scale assessments should not restrict themselves to a single mode of examination. There is no reason, save cost and complexity, that the assessment cannot be multi-method in design (i.e. that a number of techniques be incorporated into an overall assessment package). Thus, the assessment designer can utilize one or more comprehensive, multiple choice instruments in order to provide information about the knowledge base acquisition of students, and can also include one or more senior essays, problem solving simulations, oral examinations, and the like in order to assess more fully the higher order skills and proficiencies of a sample of undergraduate majors. A mix of these various formats and contents could be selected to represent the relative importance of each component in the curriculum. A multi-method formulation of an assessment would also allow the relative strengths and weaknesses of the program to be assessed in terms of cognitive skills as well as the specific content of the curriculum.

Sampling and Program Evaluation

When selecting the sample of students to be included in the assessment, it is important to recall the purpose of the assessment. If the goal of the assessment program is to measure the efficacy of the instructional program (as opposed to certifying the individual student), then while students are the means by which information on program effectiveness is obtained, they are not the ultimate focus of the assessment. Nonetheless, it is necessary that a procedure be developed such that once a comprehensive program is fully implemented (i.e., functioning in all majors), each and every undergraduate at an institution is eligible for assessment in at least one major. It is not necessary that a single strategy for sampling be adopted on an institution-wide basis, but rather, the sampling approach should reflect the nature of the particular unit of the assessment. The simplest method of obtaining the sample is simply to include all students in a given major in the assessment of that major, as would be necessary if student certification were the goal of the assessment. This approach may work in those programs that are small enough to allow such an inclusion rule, but in a majority of cases it will be necessary to use a sample.

The first step in selecting a sample is the formal specification of an inclusion rule (i.e. who is eligible to be included within a specific assessment package). The listing of all eligible students is technically called the sampling frame. Among the considerations necessary to specify the sampling frame are: Should the assessment be limited to senior majors only? Should part time students be included? Should double majors be included in the assessment if the area being assessed is their second major? Should "concentrators" (i.e. students whose formal degree programs include less than a full major in the area being assessed) be included? Should "special" (i.e. non-degree students) be included? The answer to these and related issues will depend upon the mission statement of the particular program and the goal of the assessment program. Whatever the answers might be, it is important that they be formally specified prior to the onset of any sampling plan and that the rules be consistent from unit to unit.

The second step in selecting a sample is to specify the sampling method to be followed. There are four essentially different types of samples which may be drawn--a random sample, a stratified (or quota) sample, a representative sample, and a volunteer sample. Each of the samples differ in terms of the validity of the conclusions which may be drawn from the ensuing assessment and in the difficulty of obtaining the sample. A random sample (the basic sample of most statistical theory) is a sample in which each and every unit (individual) in the sampling frame has the same probability of being included in the sample actually drawn. In order to draw a random sample, one must first have a complete listing of all eligible students (determined by the considerations discussed above) and then a random process that allows a sample of a predetermined size to be drawn from the sampling frame. Having obtained the listing of the sample, each individual so selected must then be individually contacted and persuaded to participate in the assessment.

The advantage of the random sample is its validity. Since each and every individual has the same probability of being selected, there is no long-range possibility of the sample being biased as long as all selected individuals participate in the assessment. If a random sample is actually obtained, all of the extensive power of statistical and psychometric theory can be brought to bear on the analysis of the resulting data, and the conclusions drawn can be framed in proper probabilistic terms.

There are, however, several difficulties in using the random sample that should not be overlooked. The first of these is that the sample is unbiased only in the long range sense. It is perfectly possible that a truly random sample will not "look right"--that is, it may lack face validity. In drawing a random sample, certain groups of individuals may be under-represented or not represented at all (e.g. if there are relatively few minority students in a particular major, it may be that that minority group will not be included at all in the assessment sample). The second problem with the random sample is that students are included in the sample without regard to their willingness to participate in the assessment. Depending on who participates in any or all of the assessment, results may be invalid. The significance of this limitation will depend upon the ability of the assessment team either to persuade the selected individuals to participate or the degree to which participation is required (e.g. is a condition of graduation) or is made a desireable activity (e.g. by offering significant payment for participation). While these problems are indeed real, the advantages of the random sample make it a worthy model for which to strive.

As an alternative to a simple random sample, one may select the sample using a stratification procedure. In this process, subgroups of interest (e.g. minorities, double majors, etc.) are first identified, followed by the drawing of a random sample of students within each of those identified groups. Under this sampling plan, it is not necessary that students be drawn from their respective groups in proportion to the group size in the total population. A "back weighting scheme" can be used to adjust the impact of each subgroup's responses in the final outcome result in proportion to the size of that group in the population. In this way, the special groups may be over-represented in the actual assessment, but then included in the final results in an appropriate manner. This approach may be particularly effective if the performance of one or more of the identified subgroupings is of particular interest at the time the assessment is conducted.

Two methods of sample recruitment which are probably best avoided in conducting an assessment are those which lead to the "representative sample" and the volunteer sample. A representative sample is one which is artificially constructed to match the demographic characteristics of the population as a whole but without the random selection found in a stratified sample. The volunteer sample is one in which the participants are exclusively those who have volunteered to serve in the assessment. Both of these sample types suffer from the possibility of biasing the results before the fact. The first is problematic because students who meet certain characteristics are intentionally "sought out" in order to build the representative sample. Volunteer samples are generally unacceptable because they are very likely to be composed of more able and serious students.

Norms

No single issue in the development of an assessment package is likely to prove more difficult and require more careful thought than that of establishing an appropriate set of norms (i.e., a scale against which to judge the results of the assessment). The selection of a norming approach will depend upon virtually all of the factors considered heretofore.

Two general approaches to the establishment of a norm can be taken: "criterion referencing" and norm referencing. Under a criterion-referenced system, the standard against which performance is judged is the minimal acceptable performance within an objective. Thus, if an instrument has been developed to assess a certain number of goals or objectives within a specific curriculum or major, the criterion-referenced approach would require the further step of specifying the minimum number of items within each objective which must be passed by the typical student in order to exhibit mastery of the objective. Results of assessments based upon a criterion-referenced system are then generally reported in terms of the percentage of test takers who have exhibited mastery in each of the objectives. The use of the criterion-referenced system depends on the development of an instrument which has been organized around specific objectives or domains as well as an agreement upon what constitutes mastery of a domain. This consensus is particularly critical when the format of assessment consists of free response questions, essays, or production tasks, and (as previously noted) when external examiners are involved.

The fact that a criterion-referenced system can be purely local (i.e., requires no information on how students outside the specific major in the particular institution being assessed perform) provides at once a great strength and a great weakness. On the positive side is the fact that one can use a locally developed instrument without having to concern oneself with the lack of national data; nor (if one is using a commercially available instrument) does one have to worry about the comparability of the local sample to the "norm" sample, as would be the case, for instance, if a random sample of majors took the GRE Area Examination where the "norm" group would consist largely of majors seeking admission to graduate school. Other strengths of a criterion-referenced system are that it provides detailed information by objective and immediately suggests domains in which excellence exists or in which improvement is needed. The major weakness of the system is that if it is created in a purely local context, there is no assurance that the results are not artificially inflated (or deflated) by the inclusion of only very easy (or difficult) items or items that are not specific to the curriculum under review.

The alternative (and more customary approach) is that of norm referencing. Essentially, this approach involves collection of data on a wide range of students at many institutions, with the results of the individual institutional assessment being judged in terms of the standing of its students in relation to the performance of the norming group. The norm-referenced approach is most desirable when one wishes to be able to make comparative statements. While norm-referenced tests (e.g. the SAT, ACT, or GRE) are used more than other forms, they are not necessarily the most desirable (the desirability of an instrument being judged in terms of the use to which it is to be put). The underlying validity of a norm-referenced test, for the purposes discussed in this essay, depends upon the content of the instrument, its technical characteristics, and the adequacy of the norming sample. Since the local institution will have little if anything to do with the establishment of the norm sample for a commercially available test, it is of particular importance that those responsible for the selection of the instrument pay as close attention to the norm sample as to the content of the test itself. It is essential that the norm sample provide an appropriate basis of comparison.

There is no reason before the fact to prefer one form of "norming" over another, and there are certainly cases where the distinction between the two forms of norming are not as clear cut as presented here. The basic validity of the criterion-referenced approach rests with the assessment instrument and the definition of the mastery criterion, that of the norm-referenced approach with the assessment instrument and the norming sample. The decision one takes will depend on the purpose of the assessment together with a consideration of these elements.

When components of the assessment are not in traditional testing formats and do not produce numerical indices (e.g. juried performance, exhibitions, or summative papers) norm scales are generally not available, although one could establish a criterion-referenced system. Although the lack of a norming system for this type of assessment may make the reporting of the outcomes of the assessment difficult, that should not discourage the use of alternative approaches to assessment.

Timing of the Assessment

Another issue of practical concern is timing. If the goal of the assessment is to evaluate the overall impact of the instructional program, then it will be necessary that the assessment be conducted after the student has completed all or nearly all of the course work in the major (including a senior thesis, etc. if that is part of the curriculum). This generally means during the last semester of the student's residency (generally the second semester of the senior year), a time well known to be rather difficult to obtain student cooperation.

Detailed planning of the exact calendar time of the assessment may be more important. Factors such as conflict with examinations, major campus events, universal testing schedules (such as the GREs), and vacation periods must be considered carefully in order to assure maximum participation. Further considerations such as holding the assessment during regular class hours, in the evening, or over a weekend period must be made. Each of these factors can have a great impact upon participation rates and student performance. In general, it is wise to schedule programs of this nature in such a way as to minimize conflicts with other scheduled activities even if it means a brief interruption of the normal academic calendar. If a multiple assessment method approach is taken, schedules can be created that offer alternative times to students selected to participate.

Student Participation

Any assessment effort which is based in part or in whole upon student responses will depend, in the final analysis, upon the magnitude and quality of student participation. There are a number of options, ranging from cash payments in the case of "one shot" assessment projects (treating participation as if it were an ad hoc campus job), to requiring a passing grade on the assessment for graduation (provided, of course, that the instrument is appropriate for assessment at the level of the individual student). Carrots may produce a full examination room, but do not guarantee maximum effort.

Given the need to have full participation from the sample as well as the need for students to give a good faith performance, this issue cannot be stressed strongly enough. After all the work that is necessary to plan and execute a first-rate assessment, it would be tragic to have the effort prove futile because of a lack of student participation, or a half-hearted effort on the part of students.

Alternatives to Student-Based Assessment: the Audit

All of the approaches discussed have involved assessment through student responses. That is, they have assumed that the proper approach to assessment is through measuring what the student has gained as a result of the instructional program. They assess the results of the sequence of exposure (curricular experience), acquisition (learning), and retention (of specifics and general "rules") as they exist in the students selected for the assessment at the time of the assessment.

It is not absolutely necessary, however, that an assessment be based on student responses in the after-the-fact manner implicit in the assessment schemes previously discussed. It is possible (although not without some limitations) to base an assessment upon a thorough cataloguing of the materials to which the student has been exposed and some indexing of the student's performance with that material at the time of exposure as well as other non-obtrusive means.

This assessment scheme would begin with the systematic collection of course syllabi, course lecture notes (if available), course quizzes and examinations, and details of course projects. In order to utilize the information contained in these documents in an optimal manner, it would be first necessary to have a detailed grid of curricular objectives prepared. The information available in the collected documents would then be set in the curricular grid on a course-by-course basis, thereby providing detailed information concerning the exposure of students to each objective. Information could also be obtained from these documents concerning student performance on each objective, as well as the stress placed on each objective in terms of its occurrence on quizzes, examinations, or projects. With such an assessment for each course offered in the curriculum, a composite could then be built from information available on student utilization of each course. With this procedure, a detailed picture of the manner in which curricular objectives are distributed could be constructed.

While it is unlikely that this approach to assessment would be seen as totally adequate (particularly from the perspective of accountability), it has its virtues. Foremost among these is that it allows a fairly accurate assessment of the degree to which putative curricular objectives are actually realized in the existing instructional program. Further, if student-based assessments identify areas of deficiency in learning, this analysis can be utilized in order to determine if the objective was actually presented or if it was presented only in a course not commonly taken by students in the major.

A variation on this approach was suggested by an anonymous reviewer of this essay. Essentially, the system involves the collection of materials which "define" the major (i.e., a statement of overall requirements, course syllabi, course quizzes, examinations and paper topics, assigned reading lists, and any general capstone documents such as senior theses). These materials are then reviewed by a panel of external examiners who are charged with assessing the major program based upon the submitted documentation. The utility of this approach depends, in large measure, on the charge given to the examiners, the questions asked of them, and the manner in which their responses are treated. The point, however, is that innovative alternatives to student-based systems for assessment are possible.

Special Problems in Certification

When considering the various purposes of assessment, special attention must be given to problems associated with the "certification of individuals" goal of assessment. If this approach is selected, special precautions will need to be taken--particularly with regard to the technical characteristics of the assessment instrument, the determination of a cutscore (if a "pass" level is to be established), and to appropriate before-the-fact communication to students that such a certification test will be required as part of the requirements for graduation. Among the additional technical requirements for an instrument used for the certification of individual students is a much higher reliability index for the instrument than is generally required for other purposes of assessment.

The problem of establishing a valid cutscore is an even more difficult issue--particularly given the legal problems that one may face if such a test or instrument is part of the criterion for graduation (or possibly for the awarding of honors, etc.). While the courts generally have not intruded in issues of the grade-setting behavior of individual instructors, it is not clear what their approach might be to institutionally set, single instrument policies. (Note that this is a different situation from the one in which a minimum grade point average criterion is used, for that criterion can be reached by performance in many individual courses.) Indeed, the use of an assessment instrument for the purpose of individual certification invites a host of legal issues. One might wish to consult Baldus and Cole (1980) for a sampling of the statistical and psychometric issues involved in the potentially discriminatory aspects of such an approach.

Finally, there is the bothersome issue of the need for prior notification of students. It is a general policy of many academic institutions that new requirements may not be instituted until they have been published and are known to an entering class (i.e., students already enrolled are generally "grandfathered" out of new requirements). Were this policy followed in the case of an assessment instrument employed for the purpose of certification, the instrument could not routinely be used (except on a volunteer basis) for up to five years after its inception, a delay that might not be acceptable.

Summary

While we have seen that there are many advantages to basing an assessment of the academic outcomes of higher education in the discipline, there are many thorny issues which any assessment planner (and administrator responsible for the assessment effort) must consider. These issues may be roughly divided into two broad classes--one which deals with the purposes for which the assessment is being conducted and the second which has more to do with technical issues involved with the design and conduct of an assessment. Both sets of issues are critical, but until the basic issues of purpose are clearly stated and agreed upon by all parties to the assessment effort (i.e. administration, faculty, and student) there is little hope that the technical issues can be resolved in such a way that the assessment will have the benefits envisioned.

End Notes

1. Scheduled to be available in the Fall of 1988, the Major Field Assessment Tests (MFATs) are constructed from "no longer used" item pools of the GRE Subject Area tests together with new items to keep the tests current. The two-hour exams are specifically designed for use in program evaluations, though they will be reliable at the level of the individual student. The tests will be normed on a scale different from those used on the GREs in order to preclude misinterpretation. Results will be reported by both gross and sub-test scores, as well as through profiles of specific strengths and weaknesses.

2. Five contracts were awarded in 1986 for 18-month studies to develop model indicators of student learning in the disciplines. The five winning fields were Biology, Chemistry, Computer Science, Mechanical Engineering, and Physics. Each contractor evaluated existing assessment instruments (both U.S. and foreign)

and methods against consensus models of learning goals, and developed frameworks for disciplinary faculty to assess learning more accurately in light of the discrepancies. Final reports were submitted in the Spring of 1988. For further information, contact: Office of Research, Division of Higher Education, 555 New Jersey Ave., N.W., Washington, D.C. 20208.

3. The OERI-sponsored project in Physics referred to above, however, demonstrated that the GRE Area Test in Physics demands both manipulation and application of variables.

References

Baldus, D.C. and Cole, J.W. Statistical Proof of Discrimination. Colorado Springs: Shepard's/McGraw Hill, 1980.

Banta, T.W. and Schneider, J.A. Using Locally Developed Comprehensive Exams for Majors to Assess and Improve Academic Program Quality. Knoxville, TN: Learning Research Center, Univ.of Tennessee, 1986.

Beaton, A. The NAEP 1983-84 Technical Report (NAEP Report No. 15-TR-20). Princeton, NJ: Educational Testing Service, 1987.

Burton, N. Trends in the Performance and Participation of Potential Graduate School Applicants. Princeton, NJ: Educational Testing Service, 1982. GREB #82-5.

Committee on Research in Mathematics, Science, and Technology Education, National Research Council. Interdisciplinary Research in Mathematics, Science, and Technology Education. Washington, D.C.: National Academy Press, 1987.

Committee of Vice Chancellors and Provosts. The External Examiner System for First Degree and Taught Master's Courses. London: Author, 1986.

DeVore, R. and McPeek, M. Report of the Content of Three GRE Advanced Tests. (GREB #78-4R). Princeton, N.J.: Educational Testing Service, 1985.

Dressel, P.L. Handbook of Academic Evaluation. San Francisco: Jossey-Bass, 1976.

Fong, B. The External Examiner Approach to Assessment. Washington, D.C.: Association of American Colleges, 1987.

Frederiksen, N. "The Real Test Bias: Influences of Testing on Teaching and Learning," American Psychologist, vol. 39 (1984), pp. 193-202.

Frederiksen, N. and Ward, W.G. "Measures for the Study of Creativity in Science Problem-Solving," Applied Psychological Measurement, vol. 2 (1978), pp. 1-24.

Jones, L.V. Educational Assessment as a Promising Area of Psychometric Research. Paper delivered at the annual meeting of the American Psychological Association, New York, N.Y., September, 1987.

Lawton, R. "The Role of the External Examiner," Journal of of Geography in Higher Education, vol. 10 (1986), pp. 41-51.

Levine, H., McGuire, C. and Natress, L. "Content Restructuring of Multiple Choice Items." Journal of Educational Measurement, vol. 7 (1970), pp. 63-74.

McCallum, L.W. "Use of a Senior Comprehensive Exam in Evaluation of the Psychology Major," Teaching of Psychology, vol. 10 (1983), pp. 67-69.

Morris, S.B. and Haas, L.J. "Evaluating Undergraduate Field Placements: an Empirical Approach," Teaching of Psychology, vol. 11 (1984), pp. 166-168.

Oltman, P.K. Content Representativeness of the Graduate Record Examinations Advanced Tests in Chemistry, Computer Science, and Education (GREB #81-12P). Princeton, NJ: Educational Testing Service, 1982.

Peterson, G.W. and Hayward, P.C. A Review of Measures of Summative Learning In Undergraduate Biology. Contract report (Contract 400-88-0057) to the Office of Researcn, U.S. Department of Education. Tallahassee: Florida State University, 1987.

Roose, K.D. and Anderson, C.J. A Rating of Graduate Programs. Washington, D.C.: American Council on Education, 1970.

Sawyer, T.M. "External Examiners: Separating Teaching from Grading," Engineering Education, vol. 66 (1976), pp.344-346.

Tannenbaum, A.G. "Teaching Political Science in Britain: The Final Examination System," Teaching Political Science, vol. 13 (1986), pp. 168-173.

Ward, W.C. "Measurement Research That Will Change Test Design for the Future," in Educational Testing Service, The Redesign of Testing for the 21st Century. Princeton, NJ: Author, 1986.

Williams, W.F. "The Role of the External Examiner in First Degrees," Studies in Higher Education, vol. 4 (1979), pp. 161-168.

Assessing Changes in Student Values

by Jerilee Grandy

Rarely, if ever, does higher education have as its sole purpose the intellectual growth of its students. The primary goals of a college or university generally include the mastery of professional and occupational skills, increased understanding of history, culture, and ideas, and the improvement of basic academic abilities. But somewhere in its history is an ideology --a hope, and a responsibility--that it will foster a sense of moral, humanitarian, and civic responsibility in our Nation's most able young adults.

Many secondary schools and colleges, especially religious-affiliated, traditional, and private institutions, are concerned about the effects they are having on student values. Attempts to measure those effects appear to be more common at the high school level than at the college level, perhaps because evaluation of any sort is more common in high schools. For the most part, the methodology, instrumentation, and problems unique to assessing values are identical at the college and secondary school levels. This is fortunate because the materials developed for high-school level assessment can generally be used equally well in colleges (Cline and Feldmesser, 1983; Kuhmerker, Mentkowski, and Erickson, 1980).

Before undertaking a study to assess changes in student values, it is useful to consider carefully the purposes of the assessment. A college will have to decide what to do with the findings. If the assessment indicates that students are not growing in these non-academic areas, will the college plan to institute new programs or alter old ones to bring about changes in values? Once there is a clear purpose for the assessment, the next step is to begin defining the values students should be gaining. A classic text that will be of interest to any college concerned with the nature of values and their measurement is Rokeach (1973).

Generally, somewhere in the college catalog or in the "statement of mission," a college will express its intention to provide an environment, or possibly a curriculum, that will help students to grow morally, spiritually, or in their sense of civic responsibility. The following examples are some actual quotations from college catalogs. Most of the statements are from well-known private colleges or universities, some of which are church affiliated. A few are from small city or community colleges. Large public universities seldom mention the development of values either in their mission statements or elsewhere in their catalogs.

1. "To provide general education and activities that will. . . enhance awareness of self and appreciation of their role in society. . . to lead in the civic, social, moral, and educational development of the community."

2. "To develop citizenship, leadership, and the skills necessary for constructive and productive life in the community."

3. "To stimulate men and women to grow in ethical responsibility."

4. "To gain knowledge of . . . human relationships to God."

5. "Appreciate aesthetic and creative expressions of humanity."

6. "Cultivate sensitivity to personal and social relationships, moral responsibilities and spiritual needs."

7. "To help a broad range of individuals gain such spiritual, civic, cultural, physical and practical education as will equip them to live well-rounded lives in a free society."

8. "To assist students as they develop an awareness of their own value as individuals with freedom and power to affect the world."

9. "To cultivate and impart an understanding of the Christian faith within the context of all forms of human inquiry and values."

10. "To help each student understand that she has many choices, that she may set her own goals and strive to fulfill them in a way that is satisfying to her . . . to encourage women to make responsible choices, without regard for prevailing convention."

11. "All specific University aims rest on the view that the human person is a being . . . whose obligation it is to develop all dimensions of personhood, the better to serve fellow humans lovingly and well."

12. "To inspire in them a desire to contribute to the culture and civilization in which they live and to form in them a trained capacity for the service of their country."

As we read these "value" goals, we may be struck with bewilderment or amusement. Whether these statements are grandiose, fuzzy, or meaningless, one thing is clear--they do not provide a sufficient basis upon which an institution can assess its effectiveness in achieving the goals they represent.

Deciding What to Assess

Assessing values differs in some fundamental ways from assessing basic skills or achievement in subject areas for which the content is well defined. More time must be spent on the initial steps of deciding what to assess and what instruments to use. Beyond that point, the usual assessment procedures can be applied to developing an experimental design, administering and analyzing instruments, and presenting findings.

In deciding what to assess, there are several types of questions that can help to clarify and operationalize college goals. What would count as evidence that a college has succeeded in inspiring students to contribute to their culture and civilization? What kind of behavior would suggest that students had failed to give their lives direction? How would we know if a student had gained a knowledge of his relationship with God? In what ways would a woman behave differently if she set her own goals rather than had them set for her? How could we determine which students seek the good in human life and which ones do not? How does a student demonstrate ethical responsibility, and how could we determine how many students do and how many do not?

Often we use words like "citizenship," "responsibility," "moral," and "ethical" without thinking much about what we mean. For purposes of assessment, and for program development as well, it can be useful to move away from those words and find more specific words that communicate more clearly. If a person develops "citizenship," does that mean that he always votes in national elections? Does that mean he is gainfully employed? Or does he simply have to stay out of jail?

If the development of citizenship is an institutional goal, for example, it is useful to compile a list of specific instances of citizenship. If moral development is a goal, instances of moral versus immoral behavior must be defined. Terminology must be refined. How is moral different from ethical? Can students be moral and not ethical, or ethical and not moral? How would they behave differently in each case? Is a Christian college more concerned about morality or ethics than other colleges?

In actual practice, it is useful to look at existing instruments to aid in defining values and goals. By examining actual items from a questionnaire, an assessment team is able to say, "Yes, this is what we want to measure," or "No, this is not what we mean by citizenship (or morality, or social responsibility, or awareness of God)." The sections that follow will illustrate part of this process for a hypothetical college.

Methods of Measurement

Methods of measurement generally fall into one of two broad categories: direct measures and unobtrusive measures. Direct measures include tests, questionnaires, and interviews. In the context of assessing values, they are ways of determining values by asking individuals directly. Unobtrusive measures are ones in which subjects are unaware that their values are being observed or measured.

Direct Measures

In the assessment of values, more than in the assessment of cognitive skills, the validity of instruments can be questionable. Put simply, an instrument is valid if it measures what it is being used to measure and not something else. A questionnaire to assess work values should measure work values--not reading ability or the desire to impress an employer, teacher, or evaluation team. Values are not concrete observables; they are what we call "constructs." Determining a student's sense of responsibility is not like determining his shoe size. "Sense of responsibility" is a construct. So are work values, nonconformity, risk-taking values, and altruism.

Estimates of an instrument's validity in measuring values are obtained in a number of ways. It is generally most desirable to have an instrument that correlates with or can be used to predict some dimension of behavior. If a score on a "moral responsibility" scale could predict what a person would actually do if confronted with the choice of whether or not to cheat on an important examination, the scale would have concurrent and predictive validity. If a person's values exist only in the abstract and have no relation to behavior, the values are not of much interest. Unfortunately, many values seem to operate in that way. A person may hold certain ideals, but when the chips are down, she does what is expedient and successfully rationalizes doing so. If possible, therefore, we try to obtain some behavioral indicators of the validity of the instrument.

By examining how well students' scores correlate with scores on other instruments designed to measure the same construct, we obtain indicators of "convergent validity." Similarly, if scores on one instrument are uncorrelated with scores on another instrument measuring something entirely different, we say it has "discriminant validity." The two types of validity combined are called "construct validity."

Construct validity is especially important because questionnaires, unfortunately, are susceptible to faking. If a person fakes by always marking a socially acceptable response, that person is likely to do so on other questionnaires. Thus, we may find a high correlation between scores on questionnaires measuring seemingly unrelated constructs. In that instance, we must conclude that the questionnaires have poor construct validity. In particular, they have poor discriminant validity.

There are primarily three ways to address the problem of faking on questionnaires. The first is to motivate respondents to be honest by promising anonymity, impressing them with the worthiness of the study, etc. A second way is to construct a "lie scale," that is, to include items in the questionnaire that describe undesirable behaviors or thoughts that all normal people have; thus to deny them would constitute lying. If the respondent has lied on that scale, we assume that he or she has probably lied on the rest of the questionnaire. Lie scales are common on personality inventories.

A third way to address the problem of faking is to supplement the assessment with other methods of measurement that do not rely on self-report. Some of the most creative ways to do this are with unobtrusive measures.

Unobtrusive Measures

The concept of unobtrusive measures was explicated in a now classic book by Webb et al. (1966) and has been applied in many educational evaluations. The measures are unobtrusive in the sense that people are observed in some way, or the results of their actions are observed, and they are either unaware of the observation or unaware that the observations are being used to make inferences about them.

One premise underlying unobtrusive measures is that any trait--including values, skills, and knowledge--leaves "traces." A child who is angry may not admit it, but he may destroy his friends' toys, steal, have tantrums, run away, or get even by refusing to eat. By cleverly selecting traces to observe, or by directly observing the person's behavior, we can (at least in theory) infer that person's values.

Devising unobtrusive measures can be a creative challenge. Often good measures arise from anecdotes. An anecdote alone can be a very misleading indicator of values, knowledge, or other traits. We often hear teachers say, for example, that their program for improving interpersonal skills is working well. Just last week two students helped another student to put on her boots, and Jimmy is no longer picking on Billy. What the teacher did not see (or acknowledge, even to himself or herself) was that four other students got into a fight, and a girl went home crying because two other girls excluded her from a game.

Although an instructor's anecdotes generally serve to express his or her biases concerning the effectiveness of a program, those same kinds of anecdotes can function as data if they are systematically observed and catalogued. Suppose that good citizenship and honesty are values that a college is trying to instill. How would those values be manifested? We could begin by looking at what behaviors we would like to see increased and what behaviors we would like to eliminate. Perhaps stealing textbooks is a problem. A student who has a book stolen reports it to Security. Therefore there is a record of the number of thefts. Now if the college begins a program whereby students examine and discuss their values, for example, it is possible to observe the record of stolen books to see if the rate of theft declines while the program is in effect.

The notion of using unobtrusive measures is theoretically appealing, but hard to translate into practice. The measures themselves can be difficult to design, and they are often affected by forces other than the program being assessed. In the case of stolen textbooks, the number of students reporting thefts may rise because of increased sensitivity to issues of honesty. The increase in reporting will appear as an increase in thefts. On the other hand, students responsible for textbook thefts may find out that someone is counting them and fear they will be caught. They turn to a new mode of crime that is not being observed.

When the unobtrusive measures take the form of observation, there are additional problems. Students generally know when they are being observed and can be on their best (or worst) behavior. Observers can be biased, despite the evaluator's attempt to hire impartial observers. Essentially no one can be a truly objective observer. Furthermore, there tends to be little agreement among the observations made by different observers. That means the observations have low reliability. There is also the question of whether it is ethical to observe people who do not know they are being observed (i.e., spy on them) and to pass judgment on their behavior.

While these problems may seem overwhelming, highly effective unobtrusive measures can and have been designed, and the ethical problem can be avoided (Sechrest, 1979). It is possible, for example, to observe the traces of behavior of a group of people without identifying a specific individual. An example of an unobtrusive measure that was very effective and did not point the finger at a specific person was a measure that was used in a high school, but would be equally applicable to colleges. Neighbors had complained to the school that students were going to a local fast-food restaurant for lunch and, upon their return to school, were depositing their hamburger and drink containers in their yards. In response, the school implemented a solution to the problem. To determine whether they were successful in reducing the litter problem, they counted the number of trash items dropped on a sample of ground area before and after the implementation of the program. They observed a sizable reduction. Furthermore, there were no more complaints from neighbors. (The solution to the problem, incidentally, was to put a committee of students in charge of running the cafeteria so they could have the kind of food they wanted, and, in addition, learn to manage a successful restaurant business.)

The design of successful unobtrusive measures requires, above all, creativity. In the example just cited, the measure itself was quite direct. Often the evidence that leads us to view students as irresponsible, lazy, or destructive needs only to be quantified to become a valid and useful indicator of change.

Diversification is generally the best strategy for selecting measures of program effectiveness. Simple and clever unobtrusive measures in addition to questionnaires can work quite well. If a variety of measures show consistent changes, the likelihood that the observed changes are real is greater than if only one measure showed change.

Selecting Ready-Made Questionnaires to Measure Change

An important consideration in the selection of instruments for assessment is whether students must be compared with other students, either in the Nation as a whole or in other specific colleges or geographic regions. The answer to that question determines, to a large extent, the particular instruments selected. If an institution does not need to compare the values of its students with national norms, or with the values of students from specific colleges, then it is not restricted to the use of survey instruments that have been used to establish norms. The college may invent its own or use widely available ready-made questionnaires.

There are a considerable number of ready-made instruments on the market and they must be selected with care. Once an institution has defined the values it wishes to assess, there are several routes it can take to obtain a list of instruments that may be suitable. One is to conduct a literature search of journals in education and psychology to find articles in which those values are discussed. A simple scanning of the abstracts will reveal the names of the instruments used. Then, if the instrument seems relevant, further reading of the article will provide useful evaluative information.

Evaluation of the quality of the questionnaire can begin by reading articles and progress later to reading documentation supplied by the publisher. These are eight important criteria against which the quality of the instrument should be judged:

(1) Generally, it should have good face validity. In non-cognitive assessment, however, an item-writer must take great care to avoid suggesting a right answer. Items requiring agreement or disagreement should be worded so that respondents do not feel that they are being incriminated by answering either way. Approximately half of the items should endorse the desired value, and agreement with those items should be scored positively. The remaining statements should endorse alternative values; they should be scored on a different scale, or, if they actually contradict the desired value, they may be scored negatively.

(2) Items should be written clearly, with each item stating just one idea, not two or three muddled together with "and" or "or." Items should be in a familiar and easy-to-understand format. The student should not have to spend time figuring out what is required to answer a question.

(3) Items should not suggest a racial, ethnic, or gender bias. "Person" should be used instead of "man." Stories should not put characters into stereotypic sex-roles or racial minorities into stereotypic occupations. Many values inventories were created decades ago, before public awareness had been directed to race or gender bias in tests. Before using a published instrument in which these biases are evident, it is advisable to consult with the publisher on possibilities for modifying the wording to remove obvious bias.

(4) The reading level should be slightly lower (preferably) than the reading level to which the students are accustomed. Time-consuming reading passages or stories are inappropriate if they challenge the student's reading comprehension or patience.

(5) Documentation must provide evidence of validity. It should describe the construct being measured, how the items were created, their intended use, the rationale and procedures used to validate the instrument, correlations with other pertinent variables, and analyses by sub-groups, particularly by race and gender. If norms for various occupational sub-groups are available, these may have useful implications for validity. For example, we might expect artists to value aesthetics more highly than business majors do, and business majors to value a large income more highly than artists do.

(6) If the questionnaire has sub-scales, each yielding a sub-score, the sub-scores should not be highly

correlated with one another. The correlation between sub-scales should be less than the reliability coefficients, otherwise the sub-scales lack discriminant validity--that is, they are not likely to be measuring different constructs.

(7) Reliability estimates must be provided. High test-retest reliability implies that responses have proved to be stable over time. Internal-consistency reliability (split-half, KR-20, or coefficient alpha) is quite different conceptually than test-retest reliability. High internal-consistency reliability implies that the items within the instrument are measuring very similar things. Perfect reliability would probably only occur if the instrument repeated the same item again and again. That would not be very useful. In the area of values assessment, high internal-consistency may suggest that the scale is measuring a very narrowly defined construct. Careful examination of the items will reveal whether different aspects and different manifestations of the value being measured are covered adequately. In the area of values assessment specifically, it may be desirable to use an instrument with very high test-retest reliability and moderate internal-consistency reliability. In addition, a scale or sub-scale should generally contain at least 15 or 20 items, otherwise its reliability is suspect.

(8) The publisher should provide instructions for administering the instrument as well as clear methods of scoring, interpreting, and presenting the results. These may include normative data (i.e., average scores for different groups tested), so that it is possible to interpret results in the light of other people's performance.

Not all information about questionnaires is included in journal articles, and not all existing instruments will turn up in a literature search. Another source is the Test Collection at Educational Testing Service (ETS) in Princeton. For a small fee, ETS will conduct a computer search of tests on any subject. Furthermore, preprinted searches are readily available at a nominal charge. For example, the search of instruments measuring "values" has nearly 100 entries. There is also a completed search on "moral development," some of which overlaps "values." Each entry provides the title of the test, the sub-tests, author, year, grade levels or ages for which it is appropriate, where it is available, and a short abstract. One of the advantages of consulting the Test Collection is that it contains not only published tests but unpublished ones which (although they have not undergone the scrutiny of most published tests) may be quite acceptable instruments that were used by their authors for dissertations or other special purposes.

Not all of the files in the Test Collection supply complete or up-to-date information. Before using any instrument, a college should still contact the author or publisher for additional information.

The two combined Test Collection searches mentioned above list many instruments suitable for measuring values at the college level. A sample of these are:

Altruism Scale (Sawyer)
Ethical Reasoning Inventory (Bode & Page)
Maferr Inventory of Masculine/Feminine Values (Steinmann & Fox)
Managerial Values Inventory (Reddin & Rowell)
Protestant Ethic Scale (Blood)
Quality of Life Dimensions (Flanagan)
Responsibility Test (Singh, et al.)
Risk-taking Attitudes/Values Inventory (Carney)
Student Information Form (CIRP)
Survey of Personal Values (Gordon)
Value Survey (Rokeach)
Work Values Inventory (Super)

A review of these instruments reveals that the majority do not have strong evidence of validity. Some authors, however, have presented normative data for a large of number of groups. The normative data are generally consistent with the values we might expect of those groups, and these data provide construct validity.

Nearly every instrument on this list has some limitations, but depending on how it is used and with what population, each has some strengths. While this chapter cannot possibly present a complete and thorough review of each instrument, I will examine one in detail, and then offer a few observations concerning some of the others. The purpose of this review is not to recommend or to reject any particular instrument, but rather to illustrate various aspects of a critique.

In selecting instruments from this list to review, let us consider Hypothetical College, whose mission is to "foster a spirit of service to society through productive work and responsible leadership." That college may wish to begin by reviewing the Protestant Ethic Scale, Responsibility Test, Managerial Values Inventory, and Work Values Inventory.

Not surprisingly, the Protestant Ethic Scale (Blood, 1969) was based on Max Weber's theories developed in his well-known book, The Protestant Ethic and the Spirit of Capitalism. The Scale contains eight items, four of which are pro-Protestant

Ethic and four of which are non-Protestant Ethic. An example of the first is: "A good indication of a man's worth is how well he does his job." An example of a non-Protestant ethic item is "When the workday is finished, a person should forget his job and enjoy himself." Respondents are instructed to indicate their degree of agreement with each item on a six-point scale ranging from "disagree completely" to "agree completely." Items are clearly worded, the response categories are appropriate, and the questionnaire can probably be completed easily in a minute or two.

Scores obtained on the two scales are uncorrelated. This is a fine example of the way items can be written so that half of them endorse the value system of interest (Protestant Ethic, in this case) and half of them endorse an alternative value system. The alternative-value items are stated in such a way as to be socially acceptable. The respondent is not incriminated by agreeing with them.

Because there are only four items on each scale, we would expect the internal-consistency reliabilities may be quite low. (They are not reported in the 1969 study by Blood.) On the other hand, the sub-scores do correlate with other measures, so their reliabilities must be high enough for the instrument to be useful. The pro-Protestant Ethic items have good face validity in light of Weber's theory, and research shows that scores on that scale were correlated positively with job satisfaction. On the other hand, scores on the non-Protestant Ethic scale were unrelated or negatively correlated with job satisfaction. These correlations provide fairly good construct validity for the instrument.

One fault of the instrument is its exclusively "male" language. It was validated on airmen and noncommissioned officers in the Air Force. In its existing form, it is clearly unsuitable for women. But with some editing, it could be a useful instrument for both men and women. After editing, a study of its validity for women would have to be undertaken before it could be justifiably used with a female population.

If a college were to consider using this instrument (with the necessary modified wording), it would first have to decide if the construct it measures is appropriate. Does the mission of the college include the development of the Protestant Ethic? If the mission is "service to society, productive work, and responsible leadership," is the Protestant Ethic the appropriate underlying value system? Is it the only value system? The college may be able to clarify its mission in relation to work ethics by reviewing, in detail, other available instruments.

One such instrument is the Work Values Inventory (Super, 1970). It yields scores on fifteen dimensions corresponding to sources of satisfaction in work. Some of these include altruism, management, creativity, and achievement, all of which may reflect the college mission. If Hypothetical College were to examine the forty-five items in this inventory, it may find that only about seven items are directly related to its mission statement. One of these is "add to the well-being of others." For the most part, however, the items would probably not suit the purposes of Hypothetical College.

The Managerial Values Inventory (Reddin & Rowell, 1970), by its very title, sounds as if it may be an appropriate instrument for use by Hypothetical College. This is a very unusual instrument. Each of twenty-eight items presents three statements, and the respondent is supposed to weight them in accordance with degree of agreement so that the weights total 3. This is an example of one item:

> A. Ill blows the wind that profits nobody.
> B. Genius is one percent inspiration and ninety-nine percent perspiration.
> C. Logical consequences are the scarecrows of fools and the beacons of wise men.

A serious limitation of this instrument is evident from this item. The statements are highly abstract and may demand greater reading and reasoning skills than many college students have acquired.

The Responsibility Test (Singh, et al.) attempts to assess a student's level of knowledge and judgment covering several aspects of responsibility. A typical item reads as follows:

> It is common to excuse doctors from jury duty because they
>
> a. know too many people.
> b. have a greater responsibility to their patients.
> c. are prejudiced against lower class people.
> d. legally do not have to serve.

As Hypothetical College examines this instrument further, it discovers that the test addresses issues of responsibility that go beyond responsible leadership. Because Hypothetical College has stated its mission in terms of "responsible leadership," it reconsiders that mission and decides that its mission actually goes beyond responsible leadership. The college is concerned with responsible behavior under all circumstances. As a result, Hypothetical College revises its mission statement to read "foster a spirit of service to society through productive work and acceptance of responsibility."

Further review of the instrument reveals that some items require judgment and some require purely factual knowledge. While it is important for students to acquire factual knowledge, measuring that knowledge is not part of a values assessment. Hypothetical College would like to separate these items and, for the purposes of the values assessment, eliminate the factual knowledge questions. The instrument appears not to be copyrighted, but Hypothetical College decides to contact the authors just the same and to modify the questionnaire for use in its own assessment.

This example demonstrates how one college might approach the selection of instruments. It would not stop with these four, but continue to review others for relevance to its mission. Each instrument would have to be judged in terms of the student body (its reading level, for example) and the congruence between the value as it is interpreted by the college and the value as it is represented in the questionnaire. This congruence should be apparent both at the item level and at the theoretical level. Most instruments have some base in theory, either a particular philosophy, psychological theory, or religion. That theoretical base should be examined before the instrument is used for assessment. An example of instruments with specific theoretical bases are those purporting to measure moral development. Much of the research, instruction, and publication activities related to moral development derive from the work of the late Lawrence Kohlberg, former director of the Center for Moral Education and Development program in the Graduate School of Education at Harvard University.

Kohlberg postulated that moral development follows a progression from actions motivated by punishment and obedience (Stage 1) to actions motivated by universal ethical principles (Stage 6). While moral development need not be restricted to Kohlberg's conception, the expression "moral development" typically brings to mind his name because his work is tied to an extensive body of psychological research and has a theoretical framework that includes an explicit position concerning the criteria of morality (see Kohlberg, 1979 and Kohlberg and Mayer, 1972). Kohlberg devised a "Moral Judgment Interview" in which he presented several moral dilemmas to a person and then asked questions about how he or she would deal with them. After administering these dilemmas and questions to young people, he contended that moral thought could be classified into six stages.

Soon Kohlberg's students were experimenting with moral instruction in the classroom, and as more teachers joined the "Kohlberg bandwagon," criticism naturally emerged (see Fraenkel, 1978). Controversy continues surrounding the validity of the psychological theory underlying Kohlberg's work and whether a person's stage of moral development correlates with anything of interest. Research showing gender and ethnic differences in moral development have raised debates on whether the theory and/or the instrument is biased or whether the research is faulty (Cortese, 1982; Walker, 1984; and Blake, 1985). Nevertheless, Kohlberg's work continues to be refined and revised and to remain in the forefront of values education.

One limitation of the Moral Judgment Interview is that it must be administered one-on-one and is therefore impractical for the assessment of large groups. Rest developed a more efficient multiple-choice measure of moral judgment entitled the Defining Issues Test. The DIT presents subjects with six stories representing moral dilemmas. After each dilemma, twelve issue-statements are listed and the subject is asked to indicate on a five-point scale how important each issue-statement is in making a decision about what ought to be done to resolve the dilemma. While a variety of scales can be constructed from the ratings, Rest suggests that a scale of principled reasoning (P-scale) be used. This scale consists of the weighted sum of the ratings for issues related to Stages 5 and 6 of the Kolhberg model.

Research studies have found reliabilities for the DIT around 0.8, and correlations in the 0.60s between the DIT and other instruments measuring similar constructs. Longitudinal research using the DIT has found a clear pattern of progression from lower-ordered to principled reasoning. For more information on the DIT, see Rest (1974, 1979, and 1980).

A third measure of moral development has focused on humanitarian/civic involvement issues. This measure was derived from questions on the survey designed by the Cooperative Institutional Research Program (CIRP). Results of research using CIRP data indicate that collegiate academic and social experiences are significantly related to the development of humanitarian/civic involvement values. Of these experiences, social involvement during college appeared to have the greatest positive impact on the development of these values.

Two other areas of values assessment that we have not mentioned in this chapter, and in which there are numerous research studies with associated instrumentation, are citizenship and religion. Institutions concerned with the development of citizenship may want to contact the Constitutional Rights Foundation at 6310 San Vincente Blvd., Suite 402, Los Angeles, CA 90048. This is a private, nonprofit organization that conducts research and produces curriculum materials for law and citizenship education.

For information on research and publications related to religious and values education, institutions may wish to contact the Search Institute, 122 West Franklin Avenue, Suite 525, Minneapolis, MN 55404. This nonprofit organization receives funding from Federal and State agencies, foundations, corporations, and charitable organizations for research, program development, consultation, and publications.

Selecting Instruments for Which Normative Data Exist

Each year a number of organizations conduct national surveys of college students, and many of the questions they ask pertain to values and their related behaviors. Often the instruments are in the public domain, so there are no copyright restrictions, and individual items can be chosen from the instruments if users wish to do so. Most of the questionnaires discussed above are copyrighted, so they must be purchased and should be used in their entirety.

The largest empirical study of college students in the United States is the Cooperative Institutional Research Program (CIRP) at UCLA. Sponsored by the American Council on Education, CIRP has collected and published data annually since 1966 on some 1,300 institutions and a total of over 5 million students (Astin and others, 1986). Their Student Information Form is a relatively short (four-page) questionnaire which includes about forty values items. The questionnaire asks students the degree to which they agree with a statement or the degree to which it is important to them.

Examples of some statements from the Student Information Form asking for degree of agreement are the following:

- The Federal Government is not doing enough to control environmental pollution.
- The death penalty should be abolished.
- Abortion should be legalized.
- Women should receive the same salary and opportunities for advancement as men in comparable situations.

Examples of statements for which students indicate the degree of importance of goals in their lives include:

- Influencing social values.
- Being very well off financially.
- Helping others who are in difficulty.
- Developing a meaningful philosophy of life.

Analyses of answers to these questions have shown trends over the past twenty years towards increasing materialism, concern over an uncertain economic future, and a movement away from traditional liberal arts interests into occupationally related major fields (Astin and others, 1987).

Numerous researchers have conducted studies using these data. Astin and Kent (1983), for example, analyzed gender differences as well as college characteristics associated with shifts in values. They found that in 1971, nearly twice as many men as women rated "being very well off financially" as "very important" or "essential." By 1980, the percentage of women endorsing this value had increased from 20 percent to 40 percent, while the percentage of men endorsing it rose from 39 percent to 46 percent. The gap between the sexes is clearly narrowing, while increasing numbers of both men and women regard financial well-being as highly important.

Institutions that participate in CIRP already have these data for their freshmen, as well as for the Nation as a whole, broken down by type of institution, region, and sex. These institutions may wish to survey their own students again in the senior year, using the same questions, to see how the responses have changed and to compare their students' responses with those in the CIRP follow-up data.

Another ongoing large-scale survey is entitled "Monitoring the Future." This is an annual survey of about 17,000 high school seniors from about 130 high schools, with followups for some years beyond graduation. A recent study of these data analyzing trends over the past decade found that "finding steady work" and having "a good marriage and family life" were rated among the most important values. "Finding purpose and meaning in life" was consistently rated extremely important by about 15 to 20 percent more of the women than the men (Bachman et al., 1986). Examining items and analyses from this data base may suggest instituting a similar ongoing survey of freshmen and seniors within a college to observe changes in their values during college, changes in the values of the college population (both freshmen and seniors) over time, and comparisons between the values of the students in that college and the values of comparable subgroups of students participating in the large-scale survey.

In addition to exploring national databases, colleges may wish to examine studies conducted at other institutions. While these will not provide national norms for comparison, they will present statistics for that institution and possibly others that participated. One such study, entitled "College Student Perceptions," was conducted by the State University of New York at Buffalo (Nichols, 1980). In this study, the university surveyed a sample of their incoming freshman class in 1973 and again in the spring of 1976 when they were juniors. They studied students' personal, social, intellectual, and professional development. The questions they asked focused on activities, interpersonal relationships, relationships with parents, and personal characteristics.

The study found that students experienced significant improvement in relationships with peers, parents, and others. However, by the time they were juniors, students were more pessimistic, less open to ideas, less enthusiastic and selfdisciplined, and less able to cope with success, competition, and loneliness than they were as freshmen. Whether these negative effects occur in most colleges may be worth knowing, particularly if optimism, openness to new ideas, enthusiasm, self-discipline, or ability to cope with stress happen to be traits that an institution wishes to foster in its students. An assessment battery could be designed to include a "burnout" scale.

Designing Original Questionnaires

Reviews of existing questionnaires to measure values turn up many psychometrically inadequate instruments as well as instruments inappropriate for the population being assessed or the variable being measured. After defining its goals and examining published questionnaires, a college may prefer to develop its own instrumentation or to adapt some items from existing questionnaires.

Selecting items from existing questionnaires has advantages over creating completely new items. Existing items are likely to have been adequately field tested and may be known to perform as expected. While it is possible to use a subset of items from existing instruments, several cautions are worth mentioning. First, one risk in using less than the entire questionnaire is that individual items may never have been validated, so we cannot be sure they measure the same thing as the total instrument. Moreover, they are likely to be few in number and consequently will not have as high reliability as the original instrument. Finally, if the instrument is copyrighted, it is necessary to obtain the publisher's permission to reproduce and use items. None of these cautions should discourage colleges from adapting existing instruments to meet their own needs. They should simply keep in mind that copyright laws exist to protect the publisher and that they must evaluate the psychometric characteristics of the instruments they create, just as they would if they had written the items themselves.

Designing original instruments is sometimes the best way to ensure that a questionnaire will measure exactly what the assessment team wishes it to measure and that it rests on the values held by the institution doing the assessment--not upon the values underlying a different author's work. There are countless books available providing detailed guidance in questionnaire design and item writing. Several of these are Berdie and Anderson (1974), Jacobs (1970), and Shaw and Wright (1967).

The design of a satisfactory questionnaire requires that particular attention be paid to scaling, item type, clarity of wording, and aesthetics. Before undertaking the task of designing instruments, an institution should call upon the expertise of specialists in its social science departments because there are a great many factors to take into consideration when constructing a questionnaire. Locally made instruments are often not as psychometrically sound as published instruments, but the published instruments may not be wholly adequate.

Assessment Design

Despite the fact that there is a considerable body of research on the impact of education on student values, most studies have suffered from non-rigorous experimental designs, restricted samples, and poor instrumentation. Consequently, we are limited in the inferences we can justifiably draw from those studies. Some of the research has been reviewed by Pascarella, Ethington, and Smart (1985) who have found inconsistent results (not surprisingly). While early studies by Jacob (1957) and Eddy (1959) were generally pessimistic, for example, the research of Feldman and Newcomb (1969) reached favorable conclusions. While the history of research in values education yields ambiguous findings regarding its effectiveness, it is possible to conduct a technically sound and informative study if the assessment design is rigorous, institutional goals are well defined, instruments are well selected, and the analysis is appropriate to the questions being asked.

A specific assessment design depends on a number of factors. Most important is recognizing that to assess an institution's effect on students' values requires that measurements be made at two points in time. It is not uncommon for a college to assess student values in the senior year and assume that their curriculum or environment is responsible for the development of those values. It is just as likely that the students arrived at college with the same values. Some colleges attract students already having the kinds of moral values that the college intends to foster. A Quaker college is likely to attract students who are altruistic and unconditionally opposed to war. If its seniors are more altruistic and peace loving than seniors at a large State university, the college still does not have enough information to know whether it affected those values one way or the other.

Because a college may attract students who already have the values endorsed by the college, often those students cannot "increase" in their adherence to those values. A college freshman who is already highly altruistic probably cannot become more altruistic as a result of college experiences. What can happen is "reverse maturation." Students may question the values learned in childhood, and actually "grow" in the opposite direction from that intended by the college. It is generally a goal of a college to teach students to question authority and think for themselves. Indeed, they will probably do that even if it is not an institutional goal. Thus, it should not come as a surprise that some students enter college being religious and respecting authority and leave with negative feelings towards religion and authority.

It is possible that some colleges are only interested in knowing the values of the students enrolling in their institution. Those values depend on the selection process, both in terms of which students apply and which students are admitted. If colleges are interested in the values of the students they attract, then it is reasonable to assess the values of freshmen. More likely, however, they will want to know what effect their college has had on student values. In that case, student values will be assessed at the point of matriculation and at that of graduation. The difference then may be due to institutional effect (or it could simply be maturation that would have occurred if the same students had gone somewhere else).

One way to know whether growth can be attributed to institutional effects is to select instruments for which there are national or regional norms for both freshmen and seniors. If students increase significantly in their belief that they are in control of their lives, that increase can be compared with the average change observed in the same variable for a representative national sample. If the increase is greater than the increase in the national sample, it is likely that the difference is an institutional effect.

Data Collection and Values Assessments

Regardless of whether an assessment is in a cognitive or non-cognitive area, new instruments must be pretested, psychometric properties (validity and reliability) must be studied, decisions must be made as to whether to assess an entire class or a sample, administration of the instruments must be scheduled, instructions (and possibly incentives) to students have to be considered, data analyses must be designed and programmed, and formats of reports and methods of presentation have to be considered carefully.

For the most part, the procedures involved in measuring values are no different from the procedures for measuring cognitive skills. But in a noncognitive area, especially in the measurement of values, there are features of the assessment that demand special care. Some of these are as follows:

- A pretest of a new instrument is more likely to turn up items that are ambiguous or unclear. Because words referring to values have different meanings to different people, items on values will elicit questions of the form "What do you mean by . . . ?" Colleges should be prepared to rewrite and try out values instruments more frequently than instruments measuring mathematics achievement.

- Validity will be difficult to establish. Face validity is more questionable on a test of responsibility than on a test of American history. Some behavioral measures as well as other instruments should be included in a validity study. Test-retest reliability may be low if the value being measured is easily affected by external activities. Value questions tend to be extremely vulnerable to momentary changes in feelings. If the same students respond very differently to an item at different times, that item has low test-retest reliability and is not useful because it cannot be relied upon to supply stable, consistent information.

- The schedule for the administration of the final instrument is critical. Even though apparently unreliable items may have been eliminated after the pretest, single events can still affect responses. Surveying attitudes immediately after a midterm exam, a vacation, or some other major event, is probably not a good idea. If incoming freshmen are surveyed within the first few weeks of school, it may be best to survey them as seniors at the same time of year. Quite possibly, the apparent burnout effect observed in the SUNY at Buffalo study can be attributed to the timing of the second survey. Students may be at a very low point at the end of their junior year (or the end of any year). It would be useful to administer the survey more frequently, if possible, to see if there is a time-of-year effect. Perhaps the week before graduation they will be more optimistic.

- Instructions to students must be carefully thought out and sensitively worded. A reading test may not have to be justified, but questions regarding personal issues are not generally included on examinations. Students will have to know why the questions are being asked and what will be done with the results. Who will see them? Even if confidentiality is assured, students are unlikely to answer honestly if they think the scores will reflect badly on them. A student in a conservative religious-affiliated college may not admit that she does not believe in God if she thinks her teachers, peers, or family might find out. Because the purpose of the assessment is to produce group averages, there is no need for individual identification. If possible, an assessment of values should be done anonymously, and the student's attention should be drawn to that fact.

These highlights point to some of the unique aspects of values assessment--those aspects that require special attention on the part of evaluators. While these features add to the complexity of the assessment process, colleges should not shy away from looking into this noncognitive area. An assessment of values can be done well if it is done thoughtfully, and it will yield results that institutions can use in achieving their important missions.

References

Astin, A.W., Green, K.C., Korn, W.S., and Maier, M.J. The American Freshman: National Norms for Fall, 1986. American Council on Education, Washington, D.C. and Higher Education Research Institute, University of California, Los Angeles, 1986.

Astin, A.W. et al. The American Freshman: Twenty Year Trends, 1966-1985. Cooperative Institutional Research Program. American Council on Education, Washington, D.C. and Higher Education Research Institute, University of California, Los Angeles, 1987.

Astin, H.S. and Kent, L. "Gender Roles in Transition," Journal of Higher Education, vol. 54 (1983), pp. 309-324.

Bachman, J.G., Johnston, L.D., and O'Malley, P.M. "Recent Findings from "Monitoring the Future: A Continuing Study of the Lifestyles and Values of Youth." In F.M Andrews (ed.), Research on the Quality of Life. Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan, 1986, pp. 215-234.

Berdie, D.R. and Anderson, J.F. Questionnaires: Design and Use. Metuchen, NJ: The Scarecrow Press, Inc.

Blake, C.A. "In Response to Walker's (1984) Review on Sex Differences in Moral Development." ERIC Document Reproduction Service No. ED263516, 1985.

Blood, M.R.. "Work Values and Job Satisfaction," Journal of Applied Psychology, vol. 53 (1969), pp. 456-459.

Cline, H.F. and Feldmesser, R.A.. Program Evaluation in Moral Education Princeton, NJ: Educational Testing Service, 1983.

Cortese, A.. "A Comparative Analysis of Cognition and Moral Judgment in Chicano, Black, and Anglo children." Paper presented at the Annual Meeting of the American Sociological Association, San Francisco, 1982.

Eddy, E. The College Influence on Student Character. Washington, D.C.: American Council on Education, 1959.

Feldman, K. and Newcomb, T. The Impact of College on Students. San Francisco: Jossey-Bass, 1969.

Fraenkel, J.R. "The Kohlberg Bandwagon: Some Reservations," in P. Scharf (ed.), Readings in Moral Education. Minneapolis: Winston, 1978, pp. 251-262.

Jacob, P. Changing Values in College: An Exploratory Study of the Impact of College Teaching. New York: Harper, 1957.

Jacobs, T.O. A Guide for Developing Questionnaire Items. Alexandria, VA: Human Resources Research Organization, 1970.

Kohlberg, L. "The Meaning and Measurement of Moral Development." Heinz Werner Memorial Lecture, April 1, 1979.

Kohlberg, L. and Mayer, R. "Development as the Aim of Education," Harvard Educational Review, vol. 4 (1972), pp. 449-496.

Kuhmerker, L., Mentkowski, M., and Erickson, V.L. (Eds.). Evaluating Moral Development and Evaluating Educational Programs that have a Value Dimension. Schenectady, NY: Character Research Press, 1980.

Nichols, D.L. College Student Perceptions: Three-Year Follow-up of 1973 Freshmen. A Study of Personal and Interpersonal Development. Buffalo, NY: State University of New York, 1980.

Pascarella, E.T., Ethington, C.A., and Smart, J.C. "The Influence of College on Humanitarian/Civic Involvement Values." Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., 1987.

Reddin, W.J. and Rowell, K.R. Managerial Values Inventory. Fredericton, NB, Canada: Organizational Tests Ltd., 1970.

Rest, J.R. Manual for the Defining Issues Test: An objective Test of Moral Judgment Development. Minneapolis: Author, 1974.

Rest, J.R. Development in Judging Moral Issues. Minneapolis: University of Minnesota Press, 1979.

Rest, J.R. "The Defining Issues Test: A Survey of Research Results," in Kuhmerker, Mentkowski, and Erickson (1980), pp. 113-120.

Rokeach, M. The Nature of Human Values. New York, NY: The Free Press, 1973.

Sechrest, L. (ed.) Unobtrusive Measurement Today: New Directions for Methodology in Behavioral Science. San Francisco, CA: Jossey-Bass, Inc., 1979.

Shaw, M.E. and Wright, J.M. Scales for the Measurement of Attitudes. New York, NY: McGraw-Hill Book Company, 1967.

Singh, B. et al. Development and Analysis of the Responsibility Test. Portland, ME: Portland Public Schools, no date.

Super, D.E.. Work Values Inventory. Boston, MA: Houghton Mifflin Company, 1970.

Walker, L.J.. "Sex Differences in the Development of Moral Reasoning: a Critical Review," Child Development, vol.55 (1984), pp. 677-691.

Webb, E.J., Campbell, D.T., Schwartz, R.D., and Sechrest, L. Unobtrusive Measures: Nonreactive Research in the Social Sciences. Chicago: Rand McNally & Co., 1966.

Indicators of Motivation in College Students

by Sandra Graham

Psychologists study motivation to understand why people think and behave as they do. Applied to achievement contexts like college, we would be addressing motivational concerns if we were to ask, for example: (1) Why do some undergraduates persist to degree completion despite enormous hardship, while others drop out at the slightest inconvenience? (2) What is the relationship between perceived difficulty of a particular college major and the characteristics of students choosing that major? (3) Why is it that students who are most in need of remedial help are often least likely to utilize their university's resources? (4) What accounts for the fact that some undergraduates set such unrealistic goals for themselves that failure is bound to occur? The inclusion of motivational variables in a treatment of college assessment underscores the growing importance of such questions in even the most complex models of learning and achievement. The skilled student, no matter how competent, will not perform well unless he or she is motivated.

In this essay, three dominant approaches to motivation will be examined, with particular attention to methods of assessment reflective of each approach. No attempt is made to survey the entirety of available assessments. Rather, my goal is to describe a representative sample of instruments, accessible to the college administrator, that highlight the strengths and limitations of each approach.

As the historically dominant school of thought, the personality approach to motivation will be examined first. Personality psychologists study underlying traits or dispositions within students, such as need for achievement or locus of control, that are thought to influence motivation.

A second, more contemporary perspective examined here is what we might label the "cognitive" approach to motivation. Here we are concerned not with underlying traits or dispositions, but rather with the student's cognitions (evaluations, perceptions, inferences) as determinants of achievement-related behavior. For example, how do individuals understand or think about success and failure? When college students do well or poorly on a test, what are their beliefs about the causes of these outcomes? Psychologists who study such mental events tend to be more concerned with particular experiences of success and failure, such as performance on a midterm or final exam, than with characteristics of students presumed to be stable across a wide range of situations.

A third perspective to be considered resembles a behavioral approach to motivation. Here the focus is on assessing directly observable activities of students. In experimental research, the behavioral indicators of motivation most often studied are intensity (e.g., how much effort does the student put forth?); choice (e.g., does the person elect hard or easy courses?); and persistence (e.g., does the student have a tendency to give up after experiencing failure?). In this section of the essay, we will describe representative inventories of college student activities that address some of these behavioral indicators of motivation. Thus, the focus is on what college students do motivationally rather than what they are like (personality approach) or how they think (cognitive approach).

It should be noted that the order of presenting these three motivational approaches is somewhat arbitrary. I do not intend to imply that any one approach is more credible or useful than another. But other than student demographics and achievement history, motivation is one of the most important determinants of academic success. It is therefore essential for college administrators to become acquainted with a multi-faceted approach to motivation and its assessment.

The Personality Approach to Motivation

Consider the question, "Is Lisa motivated in college?" If we wanted to take a personality approach to answering this question, we would measure those of Lisa's traits that reflect the way she customarily deals with achievement situations. The traits reflective of the motivation of college students that have been examined most frequently are need for achievement, locus of control, and anxiety.

Need for Achievement

The achievement motive is a personality trait reflecting the desire to do things well and to compete against a standard of excellence. People who are high in the achievement motive appear to be interested in excellence for its own sake rather than the rewards it brings. Given a choice between easy, moderate, or difficult tasks, individuals highly motivated to succeed find the tasks of moderate difficulty most attractive. These are the students, for example, who work very hard to excel, but might not necessarily enroll in the most advanced courses. In a similar vein, such highly motivated individuals tend to be moderate risk takers and to have career goals congruent with their abilities.

Assessment. The most widely used method to assess need for achievement is the Thematic Apperception Test (TAT) originally developed by Henry Murray and later refined for use in motivation research by David McClelland and John Atkinson (see Weiner, 1980). The TAT is a projective measure of motivation. It belongs to a family of measures like the Rorschach where individuals respond freely to ambiguous stimuli--in this case pictures of individuals engaged in some activity rather than inkblots. It is believed that the person's responses to such stimuli reveal underlying motives and desires.

The TAT is relatively easy to administer. Respondents typically meet in a group setting. Four to six TAT cards are presented sequentially on a screen, accompanied by the following directing questions: (1) What is happening? Who is the person? (2) What led up to the situation? (3) What is being thought? What is wanted? By Whom? (4) What will happen? The respondent is requested to make up a story, with four minutes of writing time allowed for each card.

Scoring a TAT entails assigning points to a written protocol based on the amount of achievement imagery present. Achievement imagery is present if a story contains reference to either: (1) a unique accomplishment, such as an invention or discovery; (2) long-term involvement concerns, such as working toward the goal of becoming a professional; or (3) competition with a standard of excellence, such as obtaining the highest score on an exam. A manual for scoring the TAT is described in Atkinson (1958), though considerable training is required to use it.

Most of the criticisms of the TAT as a measure of motivation concern questions of reliability. Its reported internal consistency--that is, the correlation between scores on individual stories--tends to be low, as is its test-retest reliability. With measures of this type, however, test-retest unreliability is not surprising inasmuch as respondents may feel compelled to write different stories on subsequent testing occasions. At a more practical level, the TAT is also criticized for the demands it makes on respondents. Writing imaginative stories about four to six pictures is extremely wearing and time consuming. Moreover, to obtain a written protocol, the respondent must be literate and fairly articulate. Indeed, there is evidence that the more verbally productive the respondent, the higher the achievement imagery score. Thus, the measurement of achievement need as a personality trait is confounded by the verbal fluency of the respondent.

Despite these limitations, the TAT remains the most widely used measure of need for achievement among college students. For example, a recent four-year study of blacks in predominantly white versus black colleges relied heavily on TAT scores in the construction of student motivational profiles (Fleming, 1984). Fleming reported that blacks who were high in the achievement motive were more satisfied with their college, had higher GPAs, and reported higher educational and professional aspirations.

There are a few other measures of the achievement motive that are more objective than the TAT. Occasionally, college administrators and researchers have relied on the achievement items in the Edwards Personal Preference Schedule (Edwards, 1959). In addition, Mehrabian (1968) developed a 34-item self-report questionnaire based on the characteristics of high need achievers. For example, an item indicating preference for intermediate difficulty is: "I would prefer a job which is important and involves a 50 percent chance of failure to a job which is somewhat important but not difficult." Students respond on 7-point scales anchored at "very strong agreement--very strong disagreement." The Work and Family Orientation Questionnaire (WOFO), in which two-thirds of the items deal with the achievement motive, taps similar characteristics (Helmreich and Spence, 1987). Respondents indicate their agreement with such statements as: "It is important for me to do my work as well as I can, even if it isn't popular with my co-workers." While these objective instruments have certain advantages in terms of test administration and scoring, none has proved to be as useful or as popular as the TAT.

Locus of Control

As a personality dimension, "locus of control" refers to stable and generalized beliefs concerning personal responsibility for outcomes. At one extreme is the internal--the individual who thinks of herself as completely responsible for her behavior and reinforcements. At the other extreme is the external--the individual who sees powerful others, luck, or circumstances beyond his control as responsible for outcomes. Internals tend to blame themselves for failure and accept praise for deserved triumphs. Externals, in contrast, neither blame themselves for failure nor view success as a result of their own efforts and abilities. Furthermore, people who are relatively internal have been shown to be more likely to exert effort to control their environment, less susceptible to social influences, better information seekers, more achievement-oriented, and better psychologically adjusted than externals (see Lefcourt, 1982). Of all the personality variables associated with motivation, locus of control is the trait that has probably been studied most extensively among college students.

Assessment. Individual differences in the tendency to perceive events as internally or externally controlled are measured by a variety of self-report scales, but the most popular of the instruments is Rotter's Internal-External (I-E) Scale (Rotter, 1966). A 29-item scale, the Rotter I-E has a forced choice format that pits an internal belief against an external belief. Some sample items and their possible responses are (underlined letters are choices indicating externality):

1. a. In the case of the well-prepared student, there is rarely ever such a thing as an unfair test.
   b. Many times exam questions tend to be so unrelated to course work that studying is really useless.

2. a. In my case getting what I want has little or nothing to do with luck.
   b. Many times we might just as well decide what to do by flipping a coin.

3. a. Without the right breaks one cannot be an effective leader.
   b. Capable people who fail to become leaders have not taken advantage of their opportunities.

The items are scored in the external direction. Thus, the higher the score the more external the individual. Six of the twenty-nine items on the scale are fillers (items designed to introduce other themes into the questionnaire so that its purpose is not so transparent). Thus the range of scores is 0 to 23.

The reliability of the I-E scale is reasonable, with coefficients around .70 for both internal consistency and test-retest reliability (Rotter, 1966). But the scale is not without its critics. One criticism has to do with its multi-dimensionality. The Rotter I-E samples a wide array of control beliefs--about school, work, politics, and interpersonal relations. These feelings may depict a number of control-related beliefs not necessarily predictive of one another. For example, belief in a politically responsive world (item 3 above) need not predict belief in a just world (item 2 above) which also may not predict belief in perceived responsibility for academic outcomes (item 1 above). Yet each of these factors certainly relates to some important aspect of perceived control. As Rotter (1975) himself indicates "[The scale] was developed not as an instrument...to allow for a very high prediction of some specific situation such as achievement or political behavior, but rather to allow for a low degree of prediction of behavior across a wide range of potential situations" (p. 62). This limitation may partly explain why the relationship between locus of control and academic achievement among college students tends to be modest (Findley and Cooper, 1983).

Anxiety

Most of us surely have experienced general uneasiness or feelings of tension in situations where the cause of such tension is not readily apparent. The term "anxiety" has come to be associated with these phenomena. A person with high trait anxiety tends to feel extremely anxious in situations perceived as threatening.

High anxiety appears to have many debilitating consequences, particularly in evaluative contexts such as college. There is a consistent negative relationship between anxiety and performance on measures of intellectual aptitude, with correlations of about -.20 reported in college populations. Furthermore, in the middle range of intelligence, where capacity is neither severely limited nor extensive, anxiety has a marked effect on achievement as measured by undergraduate GPA (Levitt, 1980). Too much anxiety also interferes with learning, particularly on complex and difficult tasks (learning simple tasks appears to be somewhat facilitated by anxiety). The current view is that the negative relationship between anxiety and complex learning is due to some form of cognitive interference that gets activated by worry about one's performance. Highly anxious individuals become so focused on their own performance and on self-deprecating thoughts (e.g., "Why am I so dumb?" "Why is this so hard for me?") that they become incapable of attending fully to the demands of the task. Psychologists tend to label this kind of uneasiness or tension associated with performance as "evaluative" or "test" anxiety.

Assessment. Although projective techniques are sometimes used, most measures of test anxiety are self-report inventories. The most popular are the Test Anxiety Scale (TAS) refined by Sarason from the original instrument developed by Mandler and Sarason (1952), and the Test Anxiety Inventory (TAI) developed by Speilberger and his colleagues (Speilberger et al., 1978).

The TAS is a 37-item true-false inventory with questions such as: "I seem to defeat myself while working on important tests," or "While taking an important examination, I perspire a lot." Typically, college males average about 17 out of a maximum score of 37 and college women average about 20, with higher scores indicating greater anxiety. The instrument has a split-half reliability of .91 and a test-retest reliability of about .82 over a six-week period.

The TAI is similar in content to the TAS. Items include:

1. The harder I work at taking a test, the more confused I get.
2. I feel my heart beating very fast during important tests.
3. I feel confident and relaxed while taking tests.

Participants respond by choosing one of the following alternatives: "almost always," "often," "sometimes," and "almost never." Average scores for college females are slightly higher than for college males. Given their similarities, it is not surprising that correlations between TAS and TAI scores are reported to be about .80.

Limitations. Self-report test anxiety measures, of which the TAS and TAI are representative, are quite popular and, because they are easy to administer and score, are widely used in college settings. Despite administrative advantages. it is important for potential users to be aware of some of the limitations of these particular measures of anxiety.

A number of the instruments use a "true-false" method of cueing responses. Such inventories are very susceptible to "response set," or the tendency of a number of individuals to choose one response category (i.e., "true") with apparent disregard for the content of items. Another problem with the measurement of any undesirable construct like anxiety is the effect of social desirability. Most of us like to think of ourselves as possessing desirable motives, feelings, and behavior patterns; we tend to deny perceived undesirable qualities like anxiety. Hence, many individuals are reluctant to endorse the items that indicate anxiety. Ensuring anonymity is one way to reduce this potential problem, as is the inclusion of some form of a "lie scale." The latter consists of self-evaluations that almost no one can deny. Some examples would be: "I do not always tell the truth" and "I sometimes get angry." The respondent who denies these behaviors is probably strongly influenced by social desirability; one would be alerted to interpret his or her anxiety score with caution.

Using Personality Measures of Motivation

Assuming reasonable validity of the instruments described in this section, in what way would they be useful to college administrators? In other words, how might one use these assessment tools for motivational change?

Of the personality approaches considered here, it is most clear what one might do with students who measure high in test anxiety. There are a number of specific interventions for dealing with test anxiety that employ a range of techniques, including relaxation training, biofeedback, and cognitive modification (see Tryon, 1980). There is considerable evidence that such interventions do work--that is, they lower the highly anxious student's self-reported worry over evaluation and, in many cases, they do have a modest impact on achievement.

Although the evidence is less clear with the other two personality measures, practitioners should consider intervention programs that help students develop the characteristics of those who are high in need for achievement and internal in locus of control. For example, assume that an incoming freshman class is given a locus of control measure, such as the Rotter I-E. From this assessment, one could then identify a group of students who are more external than their peers. Such students might then be candidates for a motivational enhancement program that might include training in self-responsibility and recognition of the contingent relationship between behaviors and outcomes. Lefcourt (1982) describes several intervention programs with college students based on the locus of control construct.

Similarly, there are numerous motivation enhancement programs designed to teach young adults the characteristics associated with the achievement motive (McClelland, 1985). These programs involve training in realistic goal setting, moderate risk taking, and accurate self-monitoring. The target population for such programs would be those students who score low in need for achievement.

Critique of the Personality Approach to Motivation

The very concept of personality assumes that there are characteristics or traits that remain stable over time. For purposes of assessment, that means we expect a personality trait to have both a great deal of predictability (e.g., internals should have higher GPAs than externals) and cross-situational generality (e.g., the high need achiever who works hard in biology should be equally motivated in the study of English literature). But critics of the personality approach, notably Walter Mischel (1973), have persuasively argued that traits are not very reliable predictors of behavior in specific situations. Much evidence suggests that the correlation between trait measures and behavior in specific situations is rarely greater than .30, which means that a full 90 percent of the variance in behavior is not explained by the trait. Furthermore, behavior in one situation tends not to be a very good predictor of behavior in other situations, with correlations again rarely exceeding .30. Thus, knowledge of personal characteristics has little of the predictability or cross-situational consistency that we expect of a good instrument.

Because they do summarize information about students, trait measures of motivation are useful and are probably best employed with other such instruments to construct profiles of student characteristics. For example, the three traits studied here tend to be correlated--the high need achiever is often low in anxiety and feels very much in control of his or her achievement outcomes. Yet when it comes to understanding the "why" of particular achievement-related behaviors, such as who actually gets good grades or who drops out after only one semester, it must be remembered that such behavior is determined by multiple factors, of which the assessed trait is only one small variable.

The Cognitive Approach to Motivation

Recall the question posed at the beginning of the prior section: "Is Lisa motivated in college?" If we follow a cognitive approach to motivation, we would be less concerned with Lisa's personality than with her achievement-related thoughts. Cognitive motivational psychologists place heavy emphasis on the role of thought as a determinant of behavior. They assume that people strive to explain and predict events--which requires constant processing of information about oneself and the environment. We turn now to this second approach to motivation where we ask: What does the motivated student think about?

Causal Attributions

One of the most important achievement-related thoughts aroused in students are causal attributions, or cognitions about why outcomes such as success or failure occur. We already encountered causal thinking to some degree in the discussion of locus of control. Recall that an internal believes that outcomes are caused by one's own actions whereas an external believes that outcomes are caused by environmental factors, including luck, fate, or powerful others. A more fine-grained analysis of causal thinking has been provided by social psychologists who look at attributions in the context of specific experiences with success and failure (Weiner, 1986). From an attributional perspective, we might want to know about the student's answers to such questions as "Why did I flunk biology?" or "Why did I get such a poor grade on my English term paper?" The answers to such "why" questions have far-reaching consequences for how students feel about themselves. Imagine the different implications for self-esteem for the student who attributes failure to poor study habits versus low aptitude. Psychologists and practitioners interested in cognitive approaches to motivation acknowledge the value in measuring what students think about themselves as guides to understanding their achievement behavior.

Assessment. Unlike the personality approach to motivation, there are far fewer standardized instruments measuring students' achievement-related thoughts. This is particularly true in the study of causal attributions for success and failure. But for those who prefer instrumentation more closely resembling standardization, two instruments have been developed for use with college students.

The Multidimensional-Multiattributional Causality Scale (MMCS) was developed by Herbert Lefcourt and his colleagues to assess specific attributions for success and failure (Lefcourt, et al., 1979). The measure consists of 48 questions, 24 dealing with achievement and 24 with affiliation. Within each domain, half of the items address failure and the other half concern attributions for success. The scale employs a four-point agree-disagree format. Participants are presented with a series of attributions for academic and related success and failure, and rate the extent to which they agree that the attribution could be a cause of their own achievement. The questions for academic failure tap the four attributions: ability (e.g., "If I were to fail a course it would probably be because I lacked skill in that area"); effort (e.g., "Poor grades inform me that I haven't worked hard enough"); context (e.g., "Often my poorer grades are in courses that the professor has failed to make interesting"); and luck (e.g., "Some of my bad grades may have been a function of being in the wrong course at the wrong time"). In validity studies on the MMCS, measures of internal consistency between items reveal reliabilities between .50 and .88 for the achievement scales and test-retest correlations ranging from .50 to .70 (Lefcourt et al., 1979).

The Academic Attributional Style Questionnaire (AASQ) is a more recent instrument, but it has already been used in a number of motivation studies involving college students (Peterson and Barrett, 1987). This is a 12-item questionnaire posing hypothetical situations of failure that range in both the degree of the failure involved and the severity of the consequences, e.g.:

1. You cannot find a book in the library.
2. You cannot get started writing a paper.
3. You cannot understand the points a lecturer makes.
4. You fail a final examination.
5. You are dropped from the university because your grades are too low.

For each of the twelve hypothetical events, the student is then asked to imagine:

> If such a situation were to happen to you, what do you feel would have caused it? While events have many causes, we want you to pick only one--the major cause of this event if it happened to you. Please write this cause in the blank provided after each event.

Thus, unlike the MMCS, students generate their own causes for failure rather than responding to a list provided by the test developer. Students then rate each self-generated cause on its underlying characteristics. That is, they indicate the extent to which the cause describes something about them or the environment (internal-external); whether it is a chronic cause of failure or something temporary (stable-unstable); and how generally this cause of failure applies to other achievement contexts as well (global-specific). The AASQ is derived from the parent Attributional Style Questionnaire (ASQ) which deals with both success and failure in diverse motivational contexts (Peterson, et al., 1982). Validation studies on the AASQ have not yet been reported, but those on the ASQ indicate a fairly reliable instrument (Peterson and Seligman, 1984).

As measured by the AASQ, certain attributional characteristics have specific motivational consequences. In one recent large-scale study, Peterson and Barrett (1987) found that college freshmen who explained bad academic outcomes as internal, stable, and global (e.g., "I have generally low intelligence" or "I'm just not interested in any of my courses") actually received lower grades during the freshman year than did students who used external, unstable, and specific causes (e.g., "I had bad teachers this semester"). Furthermore, the students with the stable and global style were less likely to make use of academic advising provided during the freshman year.

Similar results have been reported in field research by Ames and Lau (1982). They found that college students who attributed midterm failure to lack of effort were more likely to attend a specially arranged review session in advance of the next exam than were students who thought their failure was due to external causes such as tricky questions or an unmotivating instructor.

In sum, all of these studies show that causal attributions do influence achievement behavior like persistence and help-seeking. Thus, knowing something about college students' self-ascriptions for success and failure is probably quite useful in understanding why some students appear motivated and others do not, despite probable similarities in their intellectual characteristics.

Expectations and Achievement

Another achievement-related cognition that appears to be particularly important to motivation is the expectation of success. Many of the historically dominant conceptions of motivation are expectancy-value theories (Atkinson, 1964; Rotter, 1966). Motivated behavior depends on how much we value an outcome (an affective variable) and our confidence that we can achieve it (the cognitive expectancy variable). Indeed, every major cognitive motivational theorist of the twentieth century includes expectancy of goal attainment as one of the principal determinants of action (Weiner, 1986). Expectancy has been variously operationalized in terms of either a subjective probability, perceived difficulty of task (e.g., hard tasks lead to low expectations for success), perceived confidence, certainty, and/or self-efficacy.

Assessment. As with attributions, standardized instruments to measure expectations are few in number, although one often finds indirect measures of goal anticipation in many college student questionnaires. It seems quite simple to ask the college student directly how she or he expects to perform on an upcoming exam, in a particular course, or at a particular institution. Two noted standardized instruments exist for doing so.

The Non-Cognitive Questionnaire (NCQ) is an instrument developed by Tracey and Sedlacek (1984, 1987) to measure variables that theoretically should relate to academic success in college. Most of these variables are, in fact, achievement-related cognitions of the type described here. The NCQ consists of twenty-three items, comprising two forced-choice items on educational aspirations; eighteen Likert items on expectations and self-assessment; and three open-ended items assessing goals and accomplishments. The questions tapping expectations for success include such items as:

1. It should not be very hard to get a B (3.0) average at school.
2. I expect to have a harder time than most students at this school.
3. I get easily discouraged when I try to do something and it doesn't work.

Students rate these items on a 5-point scale anchored at "strongly agree-strongly disagree."

Reports on validity studies of the NCQ indicate that the instrument is psychometrically sound. Two-week test-retest reliabilities average about .85 and there is support for its construct validity when factor analysis studies have been done (Tracey and Sedlacek, 1984). In studies relating this measure to academic achievement, the NCQ is a good predictor of college GPA, typically equal to or better than predictions using SAT scores alone. One final point of interest is that the NCQ seems to be particularly predictive with minority students. The items related to self-concept and accurate self-assessment (realistic expectations) significantly predicted black student persistence as measured by enrollment figures after 3, 5, 6, and 8 semesters (Tracey and Sedlacek, 1984). A very recent comparative racial study also documented that the NCQ was predictive of persistence for those black students who graduated after 5 or 6 years of college study (Tracey and Sedlacek, 1987).

The Motivated Strategies for Learning Questionnaire (MSLQ) is a more extensive self-report developed by researchers at the University of Michigan (Pintrich, 1987). The complete MSLQ consists of 110 questions on which students rate themselves using 7-point Likert scales. Half of the items relate to information processing and meta-cognitive strategies. The remaining 55 are motivation items. Many of these address three achievement-related cognitions: expectation of success (e.g., "Compared with other students in this class, I expect to do well" or "I think the subject matter of this course is difficult to learn"); self-perceived ability ("Sometimes I have given up doing something because I thought too little of my ability" or "Compared with other students in this class, I think I have excellent study skills"); and "efficacy beliefs" ("I think my grades in this class depend on the amount of effort I exert" or "I think my grades in this class depend on the instructor's teaching and grading style"). A smaller set of motivation items relate to intrinsic interest (e.g., "I often choose course assignments that are interesting even if they don't guarantee a good grade"); and value (e.g., "I think that what I learn in this course will be useful to me after college").

In field studies with college students, the MSLQ has been used as a predictor of performance in English Composition, Biology, and Psychology courses (Pintrich, 1987). Performance measures included final course grades as well as evaluations on exams, papers, and laboratory assignments. The best predictor of performance among the cognitive measures was expectancy for success.

Critique of the Cognitive Approach to Motivation

It should also be noted that the relations described above tend to be modest, with correlations ranging from .20 to .45. Critics of cognitive approaches to motivation rely on such correlational data to caution potential users. Causal thoughts and other such mental operations are not observable, they point out, hence the test developer or user is by necessity always operating at a high level of inference. Furthermore, many critics argue that even if adequate instrumentation were available, there is still a more fundamental problem with cognition because individuals do not have direct access to their thought processes underlying judgments like causal attributions (Nisbett and Wilson, 1977). This means, for example, that a student who reports that math failure was due to low ability might not be able to articulate how she arrived at this particular attribution. To the extent that this criticism is true, it may be problematic to give cognitions such central status in efforts to assess motivation.

Finally, there is the question of the direct impact of achievement-related cognition on behavior. Cognitivists assume that behavior is a direct outgrowth of cognition, but this relationship is often hard to document. Some of these same issues arose in considering the personality approaches to motivation. As intimated in that discussion, one must be particularly careful when choosing performance as the behavioral variable of interest because exam performance, grade point average, and other such specific indications of achievement are greatly overdetermined.

The Behavioral Approach to Motivation

"Is Lisa motivated in college?" Our final approach to answering this question focuses more specifically on behavior. Here we ask: What does Lisa do that might be interpreted as a behavioral indicator of motivation? As I noted earlier, psychologists who study motivated behavior in the laboratory often look at intensity (e.g., How vigorously is the individual engaged in an activity?); choice (e.g., In what direction does the student prefer to go given a range of option?); and persistence (e.g., Does the student maintain his or her commitment to earn a college degree despite financial pressures or academic difficulties?). In this part of the essay, I focus principally on the assessment of intensity and persistence. The potentially relevant literature on choice (e.g., vocational preference, college major) is simply too disparate for this essay on assessment. Furthermore, the instruments examined in the following sections were not developed specifically as behavioral indicators of motivation. Rather, based on our conception of intensity and persistence of behavior, we infer that they tap motivational variables.

Intensity

Perhaps the clearest indicator of motivational intensity is the amount and quality of effort expended. Virtually all theories of motivation allow that one can infer motivation from how hard the individual appears to try.

A widely used instrument to measure student effort is the College Student Experiences Questionnaire (CSEQ) developed by Robert Pace and his associates at UCLA Graduate School of Education (Pace, 1984). The CSEQ is designed to measure both the level and quality of student effort in various activities of college life, including studying, reading, attending cultural events, and interacting with faculty and peers.

The instrument taps fourteen categories of experience in college life. These are mainly concerned with academic/intellectual activities (e.g., course learning, library use, writing); personal/interpersonal activities (student acquaintances, conversations); and group facilities and associations. Each category contains a list of ten to twelve activities, ranging from undemanding to very effortful, e.g., under Course Learning:

1. Took detailed notes in class.
2. Underlined major points in the readings.
3. Worked on a paper or project where you had to integrate ideas from various sources.
4. Made outlines from class notes or readings.
5. Did additional readings on topics that were introduced and discussed in class.

Students report on how often they have engaged in each of the activities during the current school year by checking "never," "occasionally," "often," or "very often." For scoring purposes, 1="never" and 4="very often." Thus, the student's score on the Course Learning Category could range from 10 for no engagement to 40 for engagement in all of the activities with great frequency.

The following are sample items from the other thirteen categories of experience tapped by the CSEQ. (1) Experiences with faculty is a category that includes ten activities ranging from routine and casual contacts (e.g., "asked your instructor for information related to a course you were taking") to more serious and long-term interactions (e.g., "worked with a faculty member on a research project"). (2) Experience in writing entails ten activities that range from "uses a dictionary or thesaurus to look up the proper meaning of words" to "revised a paper or composition two or more times before you were satisfied with it." (3) Clubs and Organizations includes ten activities ranging from simple awareness (e.g., "looked in the student newspaper for notices about campus events and student organizations") to more active involvement (e.g., "met with a faculty advisory or administrator to discuss the activities of a student organization").

Because of its recent development, the CSEQ has not been used extensively. However, Pace (1984) reports the results of a number of validation studies with the instrument that suggest sound psychometric properties. The various scales measuring quality of student effort have strong internal consistency, with reliability coefficients ranging from .79 to .90. Furthermore, the scales appear to be good predictors of student self-reported gains in intellectual skills and overall satisfaction with college (Pascarella, 1985). There is even some evidence that the scales are good predictors of who drops out and who remains in college. Pace (1984) reports a study comparing the scores on the academic quality of effort scales of community college students who transferred to UCLA and then either persisted or dropped out. Pace found that quality of effort increased overall from community college to university setting, but this difference was much greater for persisters than dropouts.

Student Involvement

A closely related construct that also appears to be a behavioral indicator of motivation is Astin's notion of Student Involvement (Astin, 1985). Astin defines involvement as "the amount of physical and psychological energy that the student devotes to the academic experience" (p. 36). Clearly, the more time and energy students invest in their college experience and the better their quality of effort, the more involved they become and the better their performance. Consistent with the perspective presented here, Astin notes the close correspondence between motivation as a psychological construct and his conception of involvement. The difference between the two is that involvement connotes behavior that can be directly observed and assessed.

According to this conception, several classes of behavior are associated with high involvement. These include (1) devoting the necessary energy to studying; (2) working at an on-campus rather than off-campus job; (3) participating actively in student organizations; and (4) interacting frequently with faculty members and student peers. Related to this last point, Astin reports that frequent interaction with faculty is more strongly predictive of college satisfaction than any other type of behavioral involvement.

Astin and his colleagues assess student involvement with data from the Cooperative Institutional Research Program (CIRP). The CIRP is an annual survey and followup of college freshmen conducted by the Higher Education Research Institute at UCLA. Over the past 20 years, about 2,300 colleges and more than 6 million college students have participated in these surveys. Participating institutions receive detailed profiles of their entering class as well as national normative data reported annually as The American Freshman.

Among the questions on the 1987 CIRP followup that tap student energy in studying are those that ask students to indicate how many hours during a given week they spent studying and attending classes. Eight-point rating scales range from "none" to "over 20 hours." Respondents are also asked how often ("not at all" to "frequently") they worked on an independent research project, stayed up all night, or failed to complete homework on time. Regarding extracurricular activities, students indicate "yes" or "no" as to whether and where they worked (on/off campus), joined any student organizations, or participated in college sports. Finally, several questions tap interactions with faculty. For example, students indicate whether they worked in a professor's home ("frequently" to "not at all"); or talked with faculty outside of class ("none" to "over 20 hours/week").

Study Skills

Another behavioral indicator of motivation implied in both the Pace and Astin questionnaires, but not directly assessed in either, is the student's study skills. Existing instruments are few but pertinent to the concerns of this essay because they generally deal with both the mechanics and conditions of studying as well as attitude toward studying and motivation to do well in academic work.

One of the most widely used measures of study skills is the Survey of Study Habits and Attitudes (SSHA) developed by Brown and Holzman (1987). For college students, the SSHA contains 100 items of which the following are illustrative:

1. I skip over the figures, graphs, and tables in a reading assignment.
2. I utilize vacant class hours for studying so as to reduce the evening's work.
3. I study three or more hours per day outside of class.
4. When preparing for an examination, I arrange facts to be learned in some logical order--order of importance, order of presentation in class, order of time in history, etc.

Respondents answer these questions on a 5-point rating scale anchored at "rarely" and "almost always".

The SSHA appears to have very sound psychometric properties. Scores on the instrument correlate well with academic achievement, with values of .42 and .45 for men and women respectively. Furthermore the internal consistency between items is very high, averaging about .88, as is the test-retest reliability. One particular advantage of the SSHA is the inclusion in the scoring materials of a Counseling Key. This allows the scorer to identify a student's responses that differ significantly from the responses commonly made by high achieving college students.

A second instrument that can be used to measure study skills in college students is the Lancaster Inventory of Approaches to Learning, developed and refined in England by Entwistle and his colleagues (Entwistle, Hanley, and Hounsell, 1979). The Inventory is a thirty-item questionnaire employing an "agree-disagree" format. Some representative items include:

1. Often when I'm reading I look out for, and learn, things which might come up on exams.
2. My habit of putting work off leaves me with far too much to do at the end of the term.
3. Distractions make it difficult for me to do much effective work in the evenings.
4. I prefer to learn the facts and details about a topic, rather than get bogged down in too much theory.

The authors of the instrument report that the items most directly related to organized study skills are moderately good predictors of first-year college grades. This is not surprising inasmuch as most of these items assess the extent to which the student works consistently, reviews regularly, and schedules work periods.

There are two similar but more elaborate instruments for the assessment of college students' study skills: the Inventory of Learning Processes (ILP) developed by Schmeck and his colleagues (Schmeck, Ribich, and Ramanaiah, 1977), and the Learning and Study Strategies Inventory (LASSI) developed by Weinstein and her colleagues (Weinstein, Schulte, and Palmer, 1987) The ILP is a 62-item "true-false" questionnaire tapping study habits with items such as:

1. I cram for exams.
2. I increase my vocabulary by building lists of new terms.
3. I generally read beyond what is assigned in class.
4. I work through practice exercises and sample problems.

Students who earn high scores on this scale purport to study more often and more carefully than other students, and the methods they claim to employ are reminiscent of what can be found in many of the old "how to study" manuals (e.g., outline the text, make up practice tests, etc.). The ILP also has good psychometric properties. Test-retest reliabilities over a two-week period range from .79 to .88. Similarly-ranged coefficients are reported for tests of internal consistency.

It is evident from the sample items listed above that all of these measures of study skills can be easily and quickly administered. However, one drawback that all of the instruments share is that more dimensions than study skills and organization are being tapped. For example, in both the ILP and Lancaster Inventory, a number of items relate to something more akin to cognitive style or preference such as, "I prefer to follow well-tried approaches to problems rather than anything too adventurous"; or "In trying to understand puzzling ideas, I let my imagination wander freely to begin with, even if I don't seem to be much nearer a solution." Furthermore, in both the ILP and the LASSI, there are a whole series of questions related to patterns or quality of information processing (e.g., "I learn new concepts by expressing them in my own words"; or "I look for reasons behind the facts"). Although there is probably a good deal of overlap between such cognitive style variables and study behavior as we have conceptualized it here, the two constructs are not the same and the potential user needs to be careful to distinguish them. Unfortunately, the literature on cognitive style assessment is too vast to be considered in this essay. The interested reader is referred to Dansereau (1985) or Weinstein and Underwood (1985) as pertinent starting points.

Persistence

Motivational psychologists view persistence at a task, particularly in the fact of failure, as a behavioral indicator of motivation. Persistence is not the same as effort, which suggests intensity of behavior. Nor is it synonymous with choice, which indicates directionality. Rather, persistence connotes continuing action toward some goal, such as obtaining a B.A. degree, when alternative activities are available, such as withdrawing from school to work full time. Not surprisingly, the construct of persistence in higher education is often associated with research on college attrition. We can therefore ask: what behaviors of college students might we measure as indicators of persistence toward degree attainment?

Aside from student background characteristics, the dominant models of college persistence appeal to a set of variables often labeled academic and social integration (Spady, 1971; Tinto, 1975). Academic integration encompasses the student's grade performance and intellectual development during college. Social integration, on the other hand, is defined as the interaction between the individual student and other persons within the college environment. Social integration occurs primarily through informal peer associations, more formal extracurricular activities, and interaction with faculty and other college personnel. Behaviorally, these constructs are most often assessed through student self-report. The fullest set of items appears to be provided in the work of Pascarella and Chapman (1983). The items associated with academic integration include: (1) hours spent studying per week; (2) number of unassigned books read for pleasure; and (3) participation in honors programs or accelerated classes. For social integration, the items tap the frequency of interaction with faculty as well as: (1) average number of weekends spent on campus; (2) number of best friends on campus; and (3) participation in informal social activities.

The reader might note the conceptual similarity between these items and those assessed by Astin's work on student involvement. One difference is that the items described above typically have not appeared on standardized instruments. Rather, researchers like Pascarella and Chapman administer questionnaires to college students at various points in their undergraduate career and include items tapping these behavioral dimensions. Student responses then yield frequency data. The higher the self-reported frequencies, the more academically and socially integrated the student is, and the more likely that student is to complete his or her college education.

Critique of the Behavioral Approach to Motivation

There appear to be a few identifiable clusters of activities in which motivated students engage. Motivated students show focused and intense study behavior, remain in college, interact frequently with faculty, and participate fully in campus life. We should probably add one important disclaimer. Samples of students say they engage in these activities. Thus, though we conceptualize the material discussed in this section as behavior, all of the assessments described rely on student self-reports. These self-reports are largely students' recollections about how frequently they engaged in a set of activities over the previous school year. That means the data are subject to memory distortions, recency effects, and other biases that threaten reliability.

To some extent, such biases will be present in any self-report. But since the focus here is on behavioral rather than psychological dimensions, this is the approach to motivation that probably would be best enriched with methods of assessment in addition to self-reports. Administrators might want to consider unobtrusive ways to observe and record student engagement in some of the motivated behaviors described in this section.

For example, many psychologists argue that observed time on task is a good indication of student effort. That means that we might want to consider noninvasive ways to measure distractibility, evidence of daydreaming, or other indications of ineffective use of time. In addition, student use of resources such as libraries or learning centers could be monitored systematically.

Regarding behaviors related to persistence, perhaps the most pressing need is for unobtrusive measures of academic and social integration. For example, we need reliable indicators of the quality as well as quantity of faculty-student interactions.

Conclusions and Recommendations

Early in this essay we asked whether a hypothetical coed named Lisa was motivated in college. To explore this issue, I posed additional questions capturing three distinct approaches to motivation. Because the focus of this essay is assessment, a number of instruments pertinent to each examined approach to motivation were described.

Given sets of student data based on these instruments, what then should be the goal of the college administrator? Where, for example, would interventions be appropriate to enhance motivation? Each of the approaches to motivation discussed in this essay offers a unique perspective on that question.

The personality approach to motivation assumes relatively stable traits in students. But traits such as need for achievement, locus of control, or anxiety proneness may not be easily modified during the college years. Therefore, college administrators might want to think about adapting the instructional experience to the characteristics of students. For example, there is evidence that highly anxious students perform better on exams when time pressures are minimized. They also show preference for courses that are highly structured. On the other hand, students high in the achievement motive often prefer situations where individual choice is maximized. Administrators might then want to ensure that their institution does indeed offer a range of courses that vary along these dimensions of structure, choice, and time constraints. To some degree, students should then be counseled to elect those courses where the method of instruction fits well with their motivational disposition.

In contrast to the presumed stability of personality traits, achievement-related cognitions of college students might be quite amenable to change. From a cognitive perspective, intervention would probably focus on counseling to alter the way students view themselves or their expectations for the future. The assumption is that a change in the way we think about success and failure will lead to a change in achievement-related behavior.

Finally, the behavioral approach to motivation can inform administrators about institutional changes that might be needed to enhance student motivation. We know that certain very specific activities of students are related to achievement behavior like persistence. These activities include getting academic help when needed, working on campus versus off, and interacting closely with faculty. Since many of these activities also reveal how students use their college facilities, they also suggest which facilities might be targeted for improvement to stimulate more usage (Pace, 1984). For example, are the resources and opportunities for academic counseling and help both adequate and easily accessible? Does the institution provide sufficient employment opportunities with competitive wages to attract students to work on campus? Does the reward structure for faculty encourage greater time and involvement with students? Does space usage within academic departments bring faculty and students into closer contact? That is, for example, do students have office space, desks, study cubicles, or lounges located in close proximity to faculty offices? Commitment to any of these kinds of broader institutional improvements can be interpreted within the framework of enhancing student motivation from a behavioral approach.

References

Astin, A. "Involvement: The Cornerstone of Excellence." Change, vol. 17, no. 4 (1985), pp. 35-39.

Atkinson, J. W. Motives in Fantasy, Action, and Society. Princeton, N.J.: Van Nostrand, 1958.

Atkinson, J. W.. An Introduction to Motivation. Princeton, N.J.: Van Nostrand, 1964.

Brown, W. F., and Holtzman, W. H. Survey of Study Habits and Attitudes Manual. New York: The Psychological Corporation, 1967.

Dansereau, D. "Learning Strategy Research." In J. Segal, S. Chipman, and R. Glaser (eds.), Thinking and Learning Skills: Relating Instruction to Research, Vol. 1, Hillsdale, N.J.: Erlbaum Publishers, 1985, pp. 209-240.

Edwards, A.L. Personal Preference Schedule Manual. New York: Psychological Corporation, 1959.

Entwistle, N., Hanley, M., and Hounsell, D. "Identifying Distinctive Approaches to Studying." Higher Education, vol. 8 (1979), pp. 365-380.

Findley, M., and Cooper, H.M. "Locus of Control and Academic Achievement: a Literature Review." Journal of Personality and Social Psychology, vol. 44 (1983), pp. 419-427.

Fleming, J. Blacks in College. San Francisco: Jossey-Bass, 1984.

Helmreich, R.L., and Spence, J.T. "Work and Family Orientation Questionnaire." Catalog of Selected Documents in Psychology, vol. 8 (1978), p.35.

Lefcourt, H. Locus of Control. Hillsdale, N.J.: Lawrence Erlbaum, 1982.

Lefcourt, H., von Baeyer, C., Ware, E., and Cox, D. "The Multidimensional-Multiattributional Causality Scale: the Development of a Goal Specific Locus of Control Scale." Canadian Journal of Behavioral Science, vol. 11 (1979), pp. 286-304.

Levitt, E. E. The Psychology of Anxiety. Hillsdale, N.J.: Lawrence Erlbaum, 1980.

Mandler, G., and Sarason, S.B. "A Study of Anxiety and Learning." Journal of Abnormal and Social Psychology, vol. 47 (1952), pp. 166-173.

McClelland, D. Human Motivation. New York: Prentice-Hall, 1985.

Mehrabian, A. "Measures of Achieving Tendency." Educational and Psychological Measurement, vol. 29 (1969), pp. 445-451.

Mischel, W. "Toward a Cognitive Social Learning Reconceptualization of Personality." Psychological Review, vol. 80 (1973), pp. 252-283.

Nisbett, R.E., and Wilson, T.D. "Telling More Than We Can Know: Verbal Reports on Mental Processes." Psychological Review, vol. 84 (1977), pp. 231-259.

Pace, C. R.. Measuring the Quality of College Student Learning Experiences. Los Angeles: Higher Education Research Institute, 1984.

Pascarella, E.T., and Chapman, D.W. "A Multiinstitutional, Path Analytic Validation of Tinto's Model of College Withdrawal." American Educational Research Journal, vol. 20 (1983), pp. 87-102.

Peterson, C.R., and Barrett, L.C. "Explanatory Style and Academic Performance Among University Freshmen." Journal of Personality and Social Psychology, vol. 53 (1987), pp.603-607.

Peterson, C.R., and Seligman, M.E.P. "Causal Explanations as a Risk Factor for Depression: Theory and Evidence." Psychological Review, vol. 91 (1984), pp. 347-374.

Pintrich, P. R. "Motivated Learning Strategies in the College Classroom." Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., April 1987.

Rotter, J.B. "Generalized Expectancies for Internal Versus External Control of Reinforcements." Psychological Monographs, vol. 80, no. 1 (1966). Whole No. 609.

Rotter, J. B. "Some Problems and Misconceptions Related to the Construct of Internal Versus External Control of Reinforcement." Journal of Consulting and Clinical Psychology, vol. 43 (1975), pp. 55-67.

Schmeck, R., Ribich, F., and Ramanaiah, N. "Development of a Self-report Inventory for Assessing Individual Differences in Learning Processes." Applied Psychological Measurement, vol. 3 (1977), pp. 413-431.

Spady, W. G. "Dropouts from Higher Education: Toward an Empirical Model." Interchange, vol. 2 (1971), pp. 38-62.

Speilberger, C.D., Gonzalez, H., Taylor, C., Algaze, B., and Anton, W. "Examination Stress and Test Anxiety." In C. Speilberger and I. Sarason (eds.), Stress and Anxiety, Vol. 5. New York: Wiley Press, 1978, pp. 44-66.

Tinto, V. "Dropout from Higher Education: a Theoretical Synthesis of Recent Research." Review of Educational Research, vol. 45 (1975), pp. 89-125.

Tracey, T.J., and Sedlacek, W. E. "Noncognitive Variables in Predicting Academic Success by Race." Measurement and Evaluation in Counseling and Development, vol. 16 (1984), pp. 171-178.

Tracey, T.J., and Sedlacek, W.E. "Prediction of College Graduation Using Noncognitive Variables by Race." Measurement and Evaluation in Counseling and Development, vol. 19 (1987), pp. 177-184.

Weiner, B. Human Motivation. New York: Holt, Rinehart, and Winston, 1980.

Weinstein, C.E., Schulte, A.C., and Palmer, D.R. Learning and Study Strategies Inventory. Clearwater, Fla.: H&H Publishing Co., Inc., 1987

Weinstein, C., and Underwood, V. "Learning Strategies: the How of Learning." In J. Segal, S. Chipman, and R. Glaser (eds.), Thinking and Learning Skills, Vol. 1. Hillsdale, N. J.: Erlbaum Publishers, pp. 241-258.

Wilson, T.D. and Linville, P.W. "Improving the Academic Performance of College Freshmen: Attribution Therapy Revisited." Journal of Personality and Social Psychology, vol. 42 (1982), pp. 367-376.

Wilson, T.D. and Linville, P.W. "Improving the Performance of College Freshmen with Attributional Techniques." Journal of Personality and Social Psychology, vol. 49 (1985), pp.287-293.

Difficulty Levels and the Selection of "General Education" Subject Examinations

by Clifford Adelman

Many of the courses that normally satisfy general education distribution requirements in U.S. colleges and community colleges are the introductory college-level courses in basic disciplines; and these, in turn, are often prerequisites for advanced work in those fields. While consensus concerning general education as a whole may be hard to find, consensus among disciplinary faculty as to the content of introductory courses is usually very high. The consensus is high enough to influence the shape of advanced work in high schools and the practice of awarding college credit on the basis of examinations such as those of the Advanced Placement Program, the International Baccalaureate, the College Level Examination Program (CLEP), and, as Woods (1985) has demonstrated with respect to two-year colleges, the College Board Achievement Tests.

These introductions to the disciplines draw large enrollments across all of American higher education. The Postsecondary Transcript Sample (PETS) of the National Longitudinal Study records some 485,000 courses taken by 12,600 students who graduated from high school in 1972 and who attended postsecondary institutions at any time between 1972 and 1984. Of those 485,000 course entries, General Psychology accounted for 2.1 percent, General Biology for 1.8 percent, General Chemistry for 1.6 percent, General Physics for 1.1 percent, and Introduction to Business Administration for 0.9 percent. Out of 1100 courses titles, then, those five accounted for 7.6 percent of all courses taken by the entire NLS/PETS sample over a 12-year period. To put it baldly, that's a lot.

Colleges assess student achievement in such topic areas as are covered by these introductions to the disciplines for one of three purposes: for granting credit-by-examination, for placement, or for course/program evaluation.

Institutions currently facing the challenge of assessing course or program effectiveness at the lower-division level often complain that there are no "appropriate" off-the-shelf instruments for doing so, and that the task of developing local instruments that can be used on a regular basis is not only conceptually formidable but also beyond their resources. To the extent to which they move forward with assessment programs in "general education" at all, they thus take refuge in such instruments as the ACT/COMP or the new ETS Academic Profile, neither of which is designed to measure grasp of disciplinary content--that is, the range of facts, assumptions, and methods that define the achievement of a novice in the field. As Centra points out, the COMP and Academic Profile are cheaper to use--by far--than batteries of achievement tests in individual subjects.

"Appropriateness": Content and Difficulty

Although there are numerous lenses through which tests are judged to be "appropriate" (including administrative time and costs), two dominate the rhetoric of this complaint. The first relates to the content validity of the available instruments in light of the local curriculum. What is "appropriate" in this first sense means what our students are likely to know given what we teach. This is a matter for content representativeness studies using retired or sample forms of the various available off-the-shelf instruments. Where a selection has been made, we presumably know what a university teaches and expects of its students.

For example, Northeast Missouri State University has publicized its very elaborate assessment program. To assess the mathematics knowledge of graduating seniors who have taken a program that certifies them to teach mathematics in secondary schools, NMSU uses the mathematics sub-test of the General Knowledge section of the National Teachers Examination (NMSU, 1984). This is a 30-minute test with 25 questions, none of which requires exposure to any subject in mathematics beyond elementary algebra and the first few lessons of plane geometry (ETS, 1976). An item content comparison with a sample of the Math/Level I Achievement Test of the College Board reveals that, in this case, virtually identical questions are asked of college seniors (on the NTE) and high school juniors (on the CEEB) concerning metrics, factors, and areas of plane figures. Likewise, a comparison of the NTE/Math section with sample SAT/Q tests reveals that virtually identical questions are asked requiring mathematical reasoning using flow charts, interpretation of graphs, and time/distance logic. When an institution makes the kind of selection that NMSU has made, it says, in effect, "this is what we teach, and this is what we expect our students to know." The NTE/Math may be very valid--hence "appropriate"--for NMSU, though one would hope that prospective high school math teachers also know trigonometry, intermediate algebra, solid geometry, elementary functions and analytic geometry, set theory, elementary statistics and probability.[1]

Examinations that cover broad content areas of general education (e.g. the CLEP General Examination in the Humanities), require a variance in the method of the typical content representativeness study before a selection is made. Analysis of the distribution of test items, by discipline, is more important than within-field item analysis. Since the CLEP examinations are used for certification purposes, test security is tight, and one is not even allowed to take notes on the specimen tests. Thus, while I took no such notes, the CLEP examination in the

Humanities struck me as over-emphasizing Art History and under-emphasizing Philosophy. A college that selects this examination for program evaluation purposes presumably offers a general education curriculum with humanities requirements that match those emphases.

The second sense of "appropriate" relates to the overall difficulty level of off-the-shelf examinations within the field, and, in relation to that difficulty level, the ability of the examinations to discriminate among degrees of student achievement. What is "appropriate," then, means how well our students are likely to demonstrate their knowledge of a particular subject on a particular instrument, and how likely the results will confirm a faculty's sense of which students are learning the material and which students are not. In other words, does the difficulty level of the examination simultaneously match the general ability level of our students and reveal degrees of achievement?

Indeed, questions concerning intended level of difficulty and distribution of difficulty levels of tasks are basic to typical test audits (ETS, 1986), and should be of no less concern to faculty. And as Graham points out in her essay, perceived difficulty is important from the student's perspective as well.

It is held that in the absence of performance data, the task of judging the difficulty levels of test items or tasks is highly problematic. At best, one must work from analogous instruments for which some performance data are available. The situation described by Hambleton and Powell (1983) in which extensive item analysis data are provided to judges for purposes of exploring "the cognitive processes involved (for the examinee population) in answering each item," and "the percentage of high and low-scoring examinees choosing each item alternative," (p. 17) is beyond the resources and capacity of most college faculty.

The technical manuals for most of the major disciplinary subject matter examinations designed for advanced, college-bound high school students and lower-division college students come with empirical data on difficulty levels (so-called "p-values") derived from the performance, by item, of the reference groups on which the examinations were normed. These data provide but an abstract notion. What we know, for example, is that on a given examination, there are 44 questions that 90 percent or more of the reference group answered correctly, 56 questions on which 50 percent of the reference group provided correct answers, and 25 questions on which only 10 percent of the group responded correctly. We usually know nothing about the reference group other than its size and the identity of institutions from which it was drawn.[2] Unless we know more about the general abilities of the students who comprised the reference group, particularly in relation to the dispersion of their scores, data from norming studies are of little help in determining the difficulty level of the examination in relation to students in a particular insitution. In determining which off-the-shelf examinations in particular subject areas are "appropriate," such data might be helpful if the same students comprised the reference group for each examination, but that simply never happens (Lenke and Beck, 1980).

It can be said, of course, that the way we have defined the difference between novices and experts is with reference to empirical observation of what they actually do when confronted with a given task, and hence, that the whole question of difficulty when experts (professors) are presenting novices (students) with tasks to elicit their knowledge and skills in a subject area is always an empirical judgment. This position, however, neglects academic conventions. That is, over time, faculty teaching introductory courses in the disciplines develop a fairly decent sense of the kind of tasks with which their students will have difficulty. Empirical observations of the past become rational touchstones of the present.

I propose that there are three methods for determining difficulty levels of "competing" examinations in a given field (and hence for at least narrowing the choice) that rely more on a combination of these touchstones and logical-rational-apriori decision rules than on aposteriori empirical analysis. I would like to "probe" each of these methods, starting with the purely rational (in the Cartesian sense), and proceeding half-way across the spectrum toward the empirical. The examinations I will use to illustrate these methods are reliable at the level of the individual student, but that should not preclude their use in program or course evaluation.

These brief explications should illustrate the ways in which faculty teaching introductory courses in a discipline can analyze an existing examination in terms of its likely difficulty for their students. In all three cases, decisions concerning content validity and difficulty are related, though (as we shall see) in different ways. By no means, though, should any faculty select an examination on the basis of these analyses alone.

Ebel's Paradigm

The first of these methods derives from Ebel's (1972) scheme for setting passing scores. It is designed for panels of judges from the same department, discipline, or course to classify questions and assessment tasks. The matrix in which each question or task must be set is one of degree of difficulty by degree of relevance to the learning objectives of the curriculum, to wit:

Relevance:	Difficulty: Easy	Medium	Hard
Essential			
Important			
Acceptable			
Questionable			

Consideration of the dimension of relevance on this matrix should occur prior to that of difficulty. After all, for purposes of judging the validity of a course examination, the difficulty of an irrelevant question is irrelevant. But once we judge a task to be "important" or "essential," Ebel's heuristic requires us to imagine student performance, not to examine student responses. That is, he asks a judge to indicate what percentage of borderline students in a course or department could answer a question correctly or complete a task in a satisfactory manner. The same question is then asked about average students, above average students, and superior students.

We might illustrate the way in which Ebel's scheme can work in the selection of instruments for assessment of discipline-related general education in the realm of history. Both the International Baccalaureate and the AP history examinations (and sometimes, the CLEP history exams in their optional essay sections) offer what is known in the history trade as the "document question." That is, students are presented with one or more related documents, and with a question that requires them to evaluate the documents as historical evidence. The question can be presented with the period and issue defined, or in a way that requires the student to identify the period and issue. If the "documents" are not texts, rather other kinds of artifacts such as photographs of buildings, city plan diagrams, cartoons, or battle plans, then the identification of period and issue is part of the objective in the first place. No matter which way the question is presented, the student must relate the documents to a particular historical context, and read them as if they were poems demanding close attention to detail and nuance. The more frames of reference required to respond to the questions, the more theory and meta-analysis involved, the more difficult a panel of faculty will judge the task.

If a faculty teaching the introductory college course in American history or world civilization has stated "mastery of synthesizing information from historical documents and artifacts" as an objective of the course, it will rate a "document question" on an examination as "essential" or "important" on Ebel's relevance scale. If a faculty regarded the mastery as "essential" and the type of question to possess a very high degree of content validity, then it might use the criterion of proportion, that is, the percentage of "document questions" in the various examinations as a first step in the selection routine. At that point, given the wording of the questions and students' familiarity with the documents used in the questions, the faculty can place each question on Ebel's matrix. Whether the faculty will select a particular off-the-shelf instrument will depend on the distribution of the questions in the matrix in light of the purpose of the assessment. A faculty, after all, may choose a "hard" test for purposes of instruction or a test of medium difficulty for purposes of certification.

Proficiency Scales

The second of the apriori rational approaches (slightly less pure than the first) derives from a well-developed methodology of assessment in foreign (or second) language education. Proficiency assessments such as those used for the language skills portions of foreign language curricula lend themselves well to rational selection because they are criterion referenced. Freed (1981) has demonstrated how a department can establish its own scales and standards in terms of nationally recognized performance criteria. The national scale is that of the Foreign Service Institute language assessments (FSI), in which there are eleven gradations between 0 and 5 for each of the four major language skills.[3]

The criteria for each FSI level of performance are described in sufficient detail for purposes of discrimination. For example, in French, the principal differences between a "2" and a "2+" on the FSI scale are:

Speaking: few grammatical errors, wider vocabulary (than the "2" level), complete control of the future tense, but still no control of the conditional or subjunctive, etc.
Listening: complete (versus "reasonably complete" for the "2" level) comprehension
Reading: can interpret questions requiring stylistic responses (emphasis mine)
Writing: more extensive vocabulary, greater facility with syntactic patterns, more frequent use of idioms.

Knowing such criteria, a department can develop its own proficiency assessments with its own scales, and decide, in advance, what level of performance will be required for different purposes. In turn, these can generate empirical comparisons. For example, in order to interpret CEEB Achievement Test scores for purposes of placing students in appropriate levels of foreign language courses, a department can establish its own reference group and match CEEB scores against the performance of the same students on its own scales which, in turn, are referenced to FSI criteria. Thus, as Freed demonstrates for an elite student population at the University of Pennsylvania, a 450 on the CEEB was equivalent to an FSI score of 1 and to a local proficiency test score of 5; a 500 was matched with 1+ and 10, and so on. Once it was determined that these scores were more or less interchangeable, they were used for purposes of placement, minimal qualifications for a grade within a course, and course evaluation.

But one must note that any system based on the FSI is not valid for assessing student performance in general education courses in a foreign language that cover literature and culture in addition to the four language skills.[4] The Advanced Placement program makes that distinction by offering two separate examinations. The CLEP and College Board Achievement tests are confined to language skills, but do not offer any free-response section, hence offer students a textual examination containing all the verbal cues that limit understanding of performance.

It is true that we do not have many FSIs (i.e. disciplines with established universal proficiency scales with descriptive ranges of performance), and hence, the bases from which to estimate the tasks a novice will find "difficult." But it is not beyond the reach of a faculty to establish local or disciplinary proficiency scales for a range of complex performance tasks, and to correlate those scales with the scales of existing measures. In other words, what the University of Pennsylvania did with foreign languages could be done by any college with General Chemistry (using the CEEB and ACS examinations), Introduction to Business Administration (using the CLEP subject exam) and others.

Within-Field Task Analysis

The third method for approaching the issue of difficulty level in test selection requires close attention to the type of cognitive operations required to answer a question or perform a task. The tradition from which this method derives may be traced to Benjamin Bloom's Taxonomy of Educational Objectives (1956), and involves analyzing learning tasks with reference to a putative hierarchy of cognitive operations. The most complex--though not necessarily the most difficult--tasks involve multiple operations at ever greater degrees of abstraction. When one adds the criteria of familiarity of subject and/or context to the task analysis, then presumably one can generate tasks that are closer to the abilities of experts, hence more "difficult" for novices.

In addition, we know that tasks are more difficult when they require a student to add a construct to the information presented (Greeno, 1978). In a way, the requirement for additional construction is a "creative thinking" prompt, for it forces a student to go beyond the givens, beyond the fixed small universe of a question, beyond the protocols of learned manipulation of givens and variables. This leap is more characteristic of the behavior of experts than it is of novices.

The assumption behind analyzing tasks in terms of a hierarchy of complexity is that it enables us to identify not merely how much a student knows, but how well a student can manipulate knowledge. It is thus an attempt to describe the quality of learning. Heywood (1977) has demonstrated the ways in which this framework can be used to formulate examinations in the disciplines, to analyze existing examinations by items, and to report out to students in terms of content domain x cognitive operation. Each question or task can be analyzed in terms of the degree to which it calls on one or more of the following:

Knowledge: of facts, terminology, conventions, principles
Comprehension: by translation (into another language, other terms, or other forms of communication); by interpretation; by extrapolation.
Manipulation: by computation, simplification, and/or solution according to single (or multiple) protocols or rules.
Analysis: by disaggregation, reorganization, distinguishing fact from assumption, matching hypotheses against facts and assumptions, inferring the relationship of part to whole.
Application: by selecting a principle of approach to an unfamiliar case, restructuring the information in the case, and application of the principle
Synthesis: by reconstruction, aggregation, judgment, evaluation.

The results of this approach are clearly reflected in both the questions and the public criteria for evaluation of International Baccalaureate (I.B.) exams. Consider an example from the 1981 International Baccalaureate in Biology. Note first--and this is important to the analysis of difficulty level--that the I.B. examinations and their accompanying curricula all make a clear distinction between "subsidiary" and "higher" levels of preparation and assessment. That these levels are based in the range of cognitive operations described above should be evident from the following. A diagram of a cell is presented, with various structures labeled:

Subsidiary	Higher
Examine the diagram of a cell [above] and answer questions 18 to 20.	Examine the diagram above which is of a cell not at division stage and answer questions 17 to 19.

18. The cell is in:
   A. interphase
   B. prophase
   C. metaphase
   D. anaphase
   E. telophase

19. Protein synthesis occurs at:
   A. H
   B. G
   C. C
   D. D
   E. A

20. The structure whose presence indicates that the cell is that of an animal and not of a plant is:
   A. F
   B. G
   C. H
   D. C
   E. D

17. At metaphase the nuclear envelope is no longer visible. On the other hand it reforms at telophase. Deduce from the figure above which of the labelled structures you think is concerned with the formation of the envelope.
A. H
B. G
C. F
D. E
E. D

18. The cell shows morphological signs of activity, such as
A. excretion
B. locomotion
C. pinocytosis
D. phagocytosis
E. none of these activities

19. [same question as #20 on the "Subsidiary" exam]

It should be obvious from a comparison of question #18/Subsidiary and #17/Higher that the former requires but knowledge/fact recall and comprehension of schematic information, whereas the latter also requires inference, hence analysis. In terms of sheer disciplinary content, the response to #17/Higher subsumes the information in #18/Subsidiary. The difference between #19/Subsidiary and #18/Higher is not only that of single versus multiple concepts, but also that of application of principle. Thus, it is, by rational analysis, a "more difficult" question, even though, by empirical analysis, some students will find the factual requirements of the questions "more difficult" than the advanced cognitive operations.

While Heywood himself expends a great deal of energy and space on empirical item analysis, my point is that the logical structure of item analysis based on a combination of content domain x complexity of cognitive operation is accessible for purposes of test selection or development to any disciplinary faculty--if that faculty is willing to spend the time.

In the courses subject to assessment in the general education portion of the U.S. college curriculum, the short answer or essay is the preferred mode of classroom assessment. When instructors grade papers or essay examinations, they tend to be more impressionistic than rigidly criterion-referenced. It is time consuming to lay out elaborate criteria--with corresponding marks--for an essay question such as those which Heywood demonstrated (criteria of the degree of organization of material, accuracy and completeness of content, and relevance and importance of analytical factors) in the case of history (Heywood, p. 48). The real issue here, though, is not the degree of detail in criteria, rather faculty consensus on what different levels of student performance actually look like, hence, the reliability of judgment. Heywood observed that when history teachers were asked to write model essay responses to questions according to content and performance criteria at four gradations, they succeeded only at the extremes.

Scales and Difficulty Levels

Having examined some rational schemes for analyzing off-the-shelf examinations, let us turn to an empirical case in the context of assessment design. Here I am interested in the way in which the notion of difficulty level inheres in locally constructed assessments through an academic convention--grading.

We customarily express our judgments of academic performance by assigning a symbol or number to an instance of student behavior. In order to do so with any credibility, we have to be able to describe the attributes of that behavior and to distinguish them from other attributes. If we engage in that discriminatory activity, we establish gradations, or benchmarks, according to the conventions of the symbolic system we have selected as a shorthand. Hence evolves a scale. The scale itself only hints at the difficulty of the questions or tasks that elicit the judged behavior of students, but empirical data derived from multiple judgments of the same performances can tell us something about the reliability of the scale.

The more crude the scale, the more reliable the judgments of student performance, or so says the conventional wisdom. With public though very general criteria, our five-grade system (A, B, C, D, F), for example, seems to work fairly well. Warren's "Academic Competences in General Education" (ACGE) examination[5], on the other hand, used as many as nine gradations per item. These gradations were determined empirically by analysis of student responses in the development phase of the examination. Each gradation for each of 47 short-answer questions in the trials of the examination carried both a faculty-generated description and model responses to assist raters. Of the 47 questions, 12 had 7 or more scoring categories, with interscorer reliability coefficients ranging from .245 to .893. While the ACGE is not a test of subject matter mastery, it illustrates the relationships between criteria, scales, and scoring that help us understand the way in which faculty think about difficulty.

In a secondary trial using 8 of the 32 questions ultimately suggested by Warren, and in a single institution with a fairly homogeneous student population of below average ability, we found the locus of interscorer reliability problems in the middle-range categories of questions with 6 or more scoring gradations. On debriefing faculty scorers, we discovered that the problem was with discrimination of performance criteria by kind, not degree. That is, in the middle-range categories, faculty felt that they were dealing as much with content domain issues as performance issues. Indeed, this has been an unstated theme of the foregoing analyses. Whether one uses Ebel's model, the proficiency model, or task analysis, content is inseparable from difficulty, as the complexity of cognitive operations is learned in the context of the disciplines.

Another issue arising from experiments with Warren's criterion-referenced short-answer examination should be noted. When presented with the selection of questions and categories of judgment prior to the assessment, faculty in the secondary trial universally agreed that one question, involving an interpretation of a brief passage from John Stuart Mill, and with five gradations in scoring (see Appendix), as generally beyond the capacities of their students, but were willing to include it. They also demonstrated the least apriori consensus on anticipated student performance with respect to an analysis of a short description of the Roman invasion of Spain, an assessment item with seven gradations of scoring (see Appendix). The interscorer reliability coefficient on the John Stuart Mill question in the development trials was .839. In the secondary trial, it was .867. In the case of the Roman invasion of Spain, the parallel results were .476 and .307 (the lowest of the eight questions we used). The point, of course, is that prior consensus on difficulty level in combination with a conventional scale (five gradations on the J. S. Mill question) has much to do with the reliability of scoring. The importance of this principle cannot be understated when faculty seek to develop alternatives to off-the-shelf instruments for assessing subject matter competence in "general education."

Summary

I intended this essay to be a "how to think about it" exercise. While empirical analyses of assessment results are helpful, they are not necessary for faculty to understand either the content validity or difficulty level of assessments in introductory college courses in the disciplines. Given instructional objectives, faculty have at least three frameworks at their disposal to analyze, apriori, course-specific examinations from the available commercial programs (CEEB, Advanced Placement, International Baccalaureate, CLEP, DANTES, etc.). On the other hand, empirical analysis of both student responses and faculty scoring is critical to ensure the reliability of judgment in "alternative," locally-developed examinations. Together, these methods are relatively efficient in helping faculty make a selection that is valid in terms of curriculum, student abilities, and the purpose of assessment.

End Notes

1. I choose mathematics because the field is least contentious, and because we have data on correct response rates, by item, for the national reference group on the NTE. Based on inspection of questions with low correct response rates, I am not confident about using superficial tables of difficulty levels drawn from empirical studies as a basis for test selection.

2. For example, the "Test Information Guide" for the CLEP General Examination in Mathematics tells us that 1,552 second-semester sophomores from 21 unnamed institutions in 1972, and 1,214 "students completing introductory level courses in mathematics" in 27 named institutions in 1978, comprised the reference groups for the test. We know nothing more about these groups (let alone what "introductory level courses in mathematics" means) except for data concerning various aspects of their actual test performance.

3. The FSI method is sometimes referred to as the LPI (Language Proficiency Interview) technique, and the scale is sometimes referred to as the ILR (Interagency Language Roundtable) scale. The Interagency Language Roundtable consists of representatives from various Federal agencies which either conduct research and/or sponsor training in second language education.

4. The ACE Commission on Educational Credit and Credentials is rather explicit about this: "No credit corresponding to the study of culture, society or literature is warranted on the basis of FSI ratings alone," and "Institutions stressing reading and writing skills will need to supplement the FSI ratings with local assessments in those areas" (Whitney and Malizio, p. 93).

5. The Academic Competences in General Education examination is copyrighted by the Educational Testing Service. The excerpts from this examination in the Appendix to this paper are reprinted with the permission of ETS.

References

Bloom, B. et al. Taxonomy of Educational Objectives. New York: David McKay, 1956.

College Entrance Examination Board. "College Level Examination Program [CLEP] Test Information Guide: General Examination in Mathematics." New York: author, 1984.

College Entrance Examination Board. "The Biology Examinations of the College Board, 1980-1982," "The History Examinations of the College Board, 1980-1982," and "The French Examinations of the College Board, 1980-1982." New York: author, 1980.

Ebel, R. The Essentials of Educational Measurement. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Educational Testing Service. A Guide to the NTE Core Battery Tests. Princeton, N.J.: author, 1976.

Educational Testing Service. ETS Standards for Quality and Fairness. Princeton, N.J.: author, 1986.

Freed, B.F. "Establishing Proficiency-Based Language Requirements." ADFL Bulletin, vol. 13, no. 2 (1981), pp. 6-12.

Hambleton, R.K. et al. Constructing and Using Criterion-Referenced Tests: AERA Training Program Materials. Amherst, Mass. Laboratory of Psychometric and Evaluative Research, 1980, photocopied materials.

Hambleton, R.K. and Powell, S. "A Framework for Viewing the Process of Standard Setting." Evaluation and the Health Professions, vol. 6, no. 1 (1983), pp. 3-24.

Heywood, J. Examining in Second Level Education. Dublin, Ireland: Association of Secondary Teachers, 1977.

International Baccalaureate Office. "Criteria of Evaluation Applicable to Examinations in 1982, 83, and 84." no date.

Lenke, J.M. and Beck, M.D. "The Ways and Means of Test Score Interpretation." In Mayo, S. T. (ed.), Interpreting Test Performance. San Francisco: Jossey-Bass, 1980, pp. 1-16.

Newble, D.I., Baxter A. and Elmslie, R.C. "A Comparison of Multiple-Choice Tests and Free-Response Tests in Examinations of Clinical Competence." Medical Education, vol. 13 (1979), pp. 263-268.

Northeast Missouri State University. In Pursuit of Degrees with Integrity. Washington, D.C.: American Association of State Colleges and Universities, 1984.

Warren, J.R. The Measurement of Academic Competence. Final Report to the Fund for the Improvement of Postsecondary Education, Grant #G007603526, 1978.

Whitney, D.S. and Malizio, A.G. Guide to Educational Credit by Examination. New York: Macmillan Publishing Co. and the American Council on Education, 1987.

Woods, J. Status of Testing Practices at Two-Year Postsecondary Institutions. Iowa City: American College Testing Service and the American Association of Community and Junior Colleges, 1985.

Appendix

Analytic Thinking Skills Questions from the Academic Competences in General Education Examination (Copyright 1977 by the Educational Testing Service, and printed here with its permission).

Definition of Analytic Competence

High: Students high in analytic skill are quick to identify the essential components of ideas, events, problems, and processes. They distinguish the important from the unimportant, fact from conjecture, and make other distinctions imaginatively but usefully. They pick out quickly deficiencies or inconsistencies in statements or positions. They are realistically skeptical or critical but not destructively so.

Middle: Middle-level students find the essence of a problem or set of ideas through methodical effort rather than quick perception. They make useful distinctions among objects or ideas but not those that are unusual or imaginative. They find the obvious deficiencies in ideas or situations they are presented and effectively follow standard or accepted procedures to identify and solve problems. They are appropriately skeptical or critical but may miss a subtle inconsistency or fail to see an unusual but available resolution of an apparent problem.

Low: Students low in analytic skill have difficulty getting beneath the surface of a problem, idea, or a situation. They work with what they are presented, rarely dissecting it to examine the source or nature of a problem. They have difficulty distinguishing the essential from the unessential or making other discriminations that would simplify their task or lead to more effective solutions to a problem. They tend to be either uncritical, accepting what they are presented without question, or blindly critical, raising questions or objections that are neither well-founded nor useful.

John Stuart Mill Quote [Question #224]

> "The only proof capable of being given that an object is visible, is that people actually see it. The only proof that a sound is audible, is that people actually hear it; and so of the other sources of our experience. In like manner, I apprehend, the sole evidence it is possible to produce that anything is desirable, is that people actually do desire it."

How could Mill's argument, and the conclusion he states in the last sentence above, best be attacked?

[The student has approximately 10 minutes to answer the question in a short written response.]

Scoring Categories:

1 The response focuses on the inadequacy of the analogy between the physical senses and a subjective response, such as desire.

2 The response points out the difference between the physical senses and a subjective response but misinterprets one or both sides of Mill's argument as in Category 3 or 4.

3 The response misinterprets both sides of Mill's argument. The student has inferred that seeing or hearing is to be the proof of existence rather than simply of visibility or audibility. The nature of desirability is also misinterpreted as, for example, requiring universal desire rather than being the prerogative of any individual, or requiring that people be aware of their desires. The response does not compare the nature of desire with the physical senses.

4 The response rests on the same kind of misinterpretation of Mill's argument as in Category 3, but deals with the physical senses or with desire but not with both and it does not contrast them.

5 The response is unintelligible, misinterprets the question, rejects the question, or misinterprets the passage so badly that the response is not pertinent.

Roman Conquest [Question # 245]

The Roman conquest of the Spanish Peninsula during the second century B.C. was far more difficult in the central plateau than in the Mediterranean coastal regions. The easily defended mountainous terrain, the increasing hostility of the natives as the Roman Legions moved further inland, and the long history of interchange with foreign peoples that made the inhabitants of the coastal regions more ready to accept Roman domination--all these contributed to the intensified resistance of the Hispanic tribes in the interior of the Peninsula.

If you were a skeptical historian, which elements in the above explanation would you accept as factual and which would you question? Explain your choices, both as to the points you accept and those you reject.

[Again, the student has approximately 10 minutes to respond.]

Scoring Categories:

1 The response accepts statements that refer to things that are observable or verifiable, such as the mountainous terrain or greater difficulty of the conquest in the interior. It questions statements that refer to differences in attitude . . . or causes of attitudes. . . .Attitudes themselves, e.g. hostility, are assumed to be verifiable. The reasons given are observability or verifiability for accepted statements, and their absence, or speculation, or the availability of alternative interpretations for questioned statements. If these reasons are clearly stated, some latitude should be allowed in the points that are accepted or rejected.

2 The response would fit Category 1, but the reasons for accepting or questioning statements are only implied or are based simply on apparent plausibility without any explanation of the reason behind the plausibility.

3 The response is mixed. Some points are justifiably accepted or rejected and others are not, or some facts may be misstated. Reasons given are at best only partially defensible, perhaps based on a sense of plausibility rather than verifiability.

4 The response is overly skeptical, rejecting almost everything that is not incontrovertibly factual and perhaps even some points that are.

5 The response draws largely from information not in the given paragraph, or adds gratuitous speculation. The added information is not just the suggestion of a plausible alternative explanation that deserves

consideration but is offered as an authoritative statement of fact.

6 The response is too short, too vaguely stated, or too limited in the information it gives to be assigned to one of the prior categories. It may focus, for example, on only one element in the paragraph or be so brief in its treatment that the information given is too sparse for any sensible judgment.

7 The response raises irrelevant issues, is incomprehensible, or does not respond to the question.

Value Added:
Using Student Gains as Yardsticks of Learning

by Leonard L. Baird

"Value added," a term frequently used to describe students' gains on tests of knowledge and skill, has attracted a great deal of attention from legislators who hope to demonstrate that public colleges are educating their undergraduates. In addition, some accrediting agencies are beginning to require colleges to produce some evidence that they have had a positive influence on the learning of their students. For example, the Southern Association of Colleges and Schools (1987) calls for institutions to " . . . ascertain periodically the change in the academic achievement of their students." A more specific approach is used by the American Assembly of Collegiate Schools of Business (AACSB). It has developed a range of instruments that permit institutions to measure subject matter competence and skill levels at various times in a student's program (Zoffer, 1987).

The growing popularity of this value added approach comes partly from the public's and legislatures' desire for evidence that students really learn enough in college to warrant the expense of higher education, and partly from colleges' own desires to evaluate their students' gains in (as opposed to absolute levels of) knowledge and skill.

Given the recency, variety of meanings and uncertainty in the value added approach, college officials and faculty face a difficult task when asked to respond to the idea. This essay is designed to assist the educator or administrator by discussing the basic concept, by explicating some of the issues surrounding value added, and by making practical recommendations to deal with the multiple questions involved. In addition to definitional problems, value added involves issues of measurement, statistical design, practicality, and basic conceptions of education.

What is Value Added? The concept originated in economics and refers to the value added at each stage of the processing of a raw material or the production and distribution of a commodity. It is usually calculated as the difference between the cost of raw materials, energy, etc. used to produce a product and the price of the product (Greenwald, 1983). This general idea is now being applied to higher education, with the "product" being the students and the "value" being the knowledge and skill students possess. The value (or knowledge and skills of students at the beginning of their college education) and their knowledge and skills at a later time are compared. The difference is held to be the value added by higher education.

However, this simple economic analogy makes many academics nervous. They object to the notion of thinking of students as products, and are particularly skeptical about attempts to quantify educational growth. As Fincher (1985) notes, they are not disposed even to think in such terms: " . . . their values, preferences and incentives have all been channeled in other directions." They are also given pause by the stated or implicit rationale that value-added information is basically for external groups that want to hold the institution accountable rather than for the internal improvement of teaching and learning. Can value added be used in a more constructive way? The answer to this question lies in its definition and the way this definition is implemented.

In practice, the meaning of value added varies considerably, from the general, such as "the institution's ability to affect its students favorably, to make a difference in their intellectual and personal development" (Astin, 1982), to rather specific blueprints that specify tests and the scores students need to obtain (Northeast Missouri State University, 1984). However, the common thread throughout these definitions is the idea of "gain," which is usually assessed by administering a test at one point, typically as freshmen enter college, and the same or a related criterion test at a later point. Although value added could refer to any criterion of interest, including personality, values, or creativity, it most commonly refers to academic achievement (e.g., increased scores on a test of mathematics or better scores than expected on knowledge of a field, such as nursing). This latter and most popular conception leads to several kinds of issues.

The Maze of Measurement. Value added assumes that we can measure change. The measurement of change is a very tricky and difficult issue, involving problems of both measurement and statistical design (Goldstein, 1983; Wood, 1986). The first problem is to find measures that will be reliable for both the initial measurement and the followup measurement, but which will yet be sensitive to students' educational growth (Carver, 1974; Kessler and Greenberg, 1981). Tests need to be reliable to provide accurate estimates of students' knowledge and capabilities. That is, a measure should give approximately the same estimate of a student's level of academic skills from one day to the next. However, the key is that the measure must also be sensitive to gain and to the influence of a college on the magnitude of gain. For example, students' height is an extremely reliable measure. To use gain in height as a criterion to assess the "value added" of a college's food services or health services would be absurd, even though height can be measured reliably and its measurement is sensitive to gain.

This hypothetical example illustrates one of the dilemmas in measurement of student characteristics. The more tests assess general characteristics, the less sensitive they are to change due to educational programs. That is, the tests become so general as to assess relatively stable characteristics of students. In the cognitive area, the more general tests border on measures of general intelligence. For example, the Watson-Glaser Critical Thinking Appraisal correlates with some measures of intelligence as highly as the two forms of the test correlate with each other, which leads to the possibility that it is measuring intelligence rather than something distinct called "critical thinking." This possibility helps to explain why most program evaluation studies that have used the Watson-Glaser have shown no effect (McMillan, 1987).

The challenge, then, is how to use or develop measures that are specific enough to be sensitive to students' educational gain but general enough to rise above the trivial. Note, too, that the tests need to be reliable at each point they are administered.

Another measurement consideration is that tests given at different points in a student's program must measure the same construct or variable (Nuttall, 1986). For example, a test of "numerical reasoning" given to freshman mathematics students might measure the capacity to think through problems, but when given to seniors, might measure memory for various algorithms. Evidence from diverse types of studies (e.g., comparisons of factor structures, correlations with other tests, etc.) can assist us in judging whether tests measure the same construct. Obviously, this task becomes problematic when the initial test is different from the later test.

The measures should also reflect the goals and specific interests of the education program. The tests need to provide information that has a fairly direct bearing on local educational practice. For this reason, off-the-shelf tests prepared by commercial testing agencies may be quite inappropriate, or at the least miss much of what the program considers important. Institutions and programs should give strong consideration to developing their own assessments. Although some institutions do not have the expertise to develop measures tailored to their own programs, and would have to call in outside experts, most probably can.

The Vicissitudes of Change: Statistical Design Issues

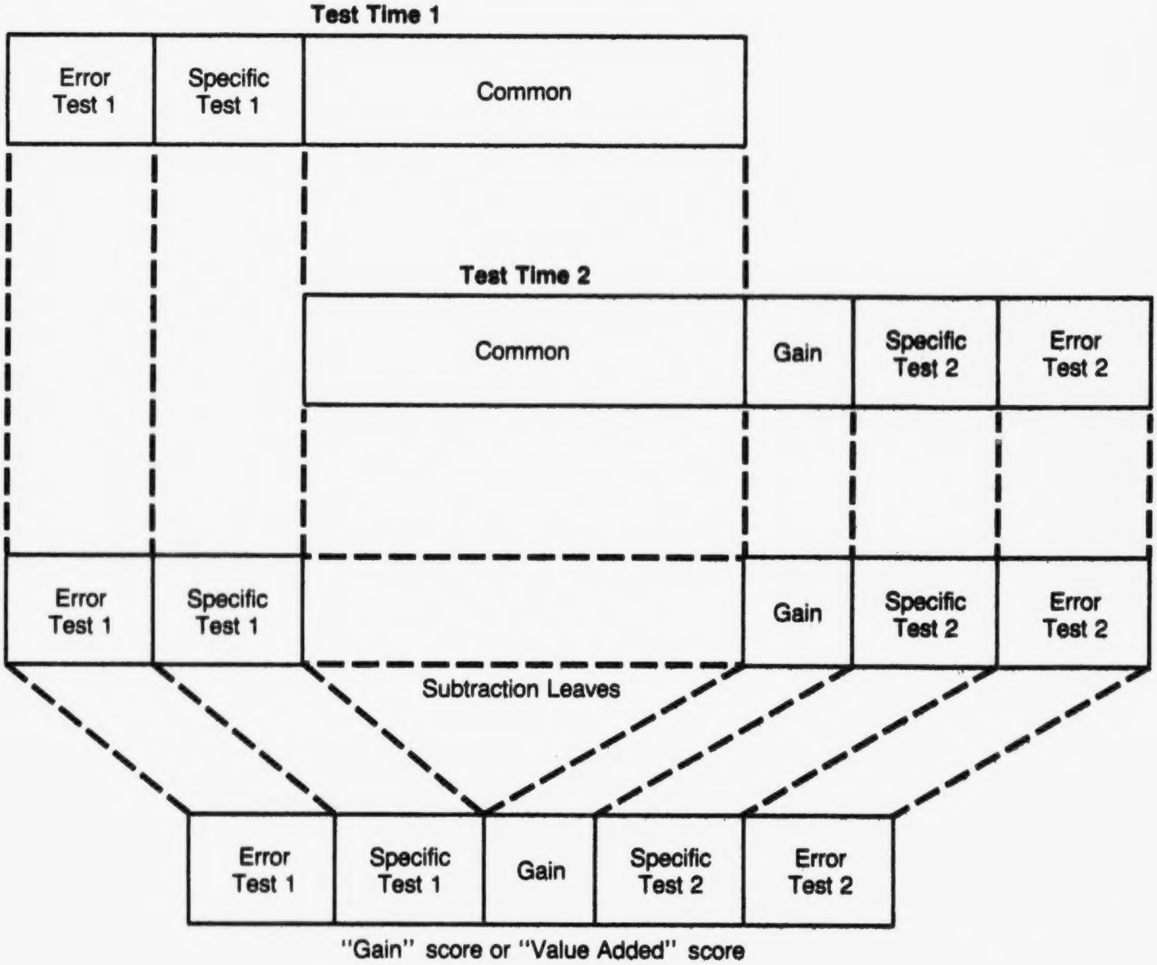
The statistical issues involved in assessing change have been debated for many years (e.g., Harris, 1963; Nuttall, 1986). Recommendations have ranged from simple change scores (final score minus initial score) to the use of latent trait models (see Traub and Wolfe, 1981). The problem, as described by Fincher (1985), is that the initial and the later test scores include both random error and test specific variance, as well as a shared variance. In other words, the two tests measure both the feature common to both tests, and different features unique to each test. The test scores also reflect the random, unsystematic aspects of students' responses to the test. As shown in Figure 1, when difference scores are used in value-added studies, part of the common variance is subtracted out, leaving a large share of the difference to be determined by the unique variances and error, as well as real change in the student.

For example, let us say that students take a vocabulary test at college entrance. Then, at the end of their freshman year, they take an alternate form of the same test. The scores on the first test are subtracted from those on the second test to provide a measure of "gain." Consider that any test score includes error that is unrelated to the variable being assessed (e.g., some items may be rather poorly written, the typography may be hard to read, testing conditions may be less than ideal, some students may be tired, distracted or unmotivated, etc.). All of these add to the randomness of the score. Likewise, each test samples a part of behavior or knowledge, so that its coverage of the content area in question will vary. Therefore, scores will vary from the first test to the second, simply because each test is, in fact, somewhat different. When the scores on the first test are subtracted from the second to yield a "gain" score, what is left over is partly due to real changes in students' vocabulary, partly due to expected errors in the tests, and partly due to the fact that the tests are not, in fact, exactly the same. Thus, it is hard to know how much faith to place in the change score.

Residual Scores

These difficulties of interpretation are increased when the first and second tests are different. For example, a college might use the SAT verbal score taken by students applying to college in their junior or senior year of high school as the initial test, and the ACT-COMP total score taken at the end of students' college sophomore year as the second test. Since the tests are clearly different, a frequently used method is to use the correlation between the two tests to produce an expected or predicted score for the second test, based on performance on the first test. For example, using hypothetical data, and based on a hypothetical correlation, a student scoring 450 on the SAT verbal might be predicted to obtain a score of 150 on the ACT COMP; a student scoring 600 on the SAT verbal might be expected to score 170 on the ACT COMP. The predicted score for students is subtracted from the actual score. The difference is known as the "residual score." If the student scores higher than predicted, the residual is positive; if the student scores lower than predicted, the residual is negative. The average residuals for students with different curricular experiences could be compared to see if students with those experiences do better or worse

Figure 1. The Nature of a Difference Score


Test Time 1
Error Test 1
Specific Test 1
Common
Test Time 2
Common
Gain
Specific Test 2
Error Test 2
Error Test 1
Specific Test 1
Gain
Specific Test 2
Error Test 2
Subtraction Leaves
Error Test 1
Specific Test 1
Gain
Specific Test 2
Error Test 2
"Gain" score or "Value Added" score

213

than expected. For example, the average residual scores of students majoring in natural sciences could be compared with those of students in the social sciences. Any differences in residual scores would be considered differences in the value added by the curriculum of those academic divisions.

The potential problems of simple change scores are compounded when this procedure is employed. The tests are different; and their unique characteristics are even more important than in the case of simple change scores. The common variance is less, and their specific variances considerably larger. Predictions based on the residual method are also imperfect; even a correlation of .60 will yield many predictions that are far off the mark. The average residual, then, can be questioned in terms of reliability--that is, there is a possibility the residual could be due to chance. Residual scores have also been criticized because they are not measures of change per se, and because performance is judged on deviations from an average prediction. Therefore, as many students will usually be below average as above, and if one program appears to be more effective, another will appear to be less effective, even if both are doing a decent job.

An additional consideration in analyses of change is that test scores can be influenced by factors other than performance on the variables of concern. For example, Astin and Panos (1969) found that performance on tests of mastery of the humanities, natural sciences, and social sciences was predicted best by the National Merit Scholarship Qualifying Test, but was also predicted by background characteristics (e.g., gender, initial career choice, parents' social class, etc.). The question for those analyzing predicted versus actual scores is whether to use all the variables that predict performance on the second test. To do so would no doubt increase the power of prediction. But some of those variables do not help us explain students' real gains; and others might be politically too delicate for a college to include in its analyses. For these reasons, difference scores are frequently criticized for their lack of consistency, stability or reliability. The solution that is often used employs the statistical techniques of regression or structural equations, but there is no clear agreement on the best ways to carry out such technical analyses. Furthermore, it is often difficult to relate results to institutional policy.

Ceiling Effects and Dropouts

Another troublesome conceptual and technical issue is the lack of independence of students' performance on the first test and the later tests. There is sometimes a correlation between the initial test performance and the gain between the first and later test. If this correlation is positive, it suggests that students with higher scores are gaining more than students with lower scores. If it is negative, it suggests that lower-scoring students gain more than higher-scoring students. Either result may be disturbing to an institution's faculty and officials. The former raises the possibility that the institution's programs are short-changing its less able students, who may not score high for a variety of personal and social reasons. The latter raises the possibility that the institution is short-changing its more able students, a considerable concern in times when institutions are searching for quality. This latter possibility is related to an artifact of many assessment designs: that when a criterion test is relatively easy, as is often the case, there is a "ceiling effect," which limits the amount of "growth" that can be shown. Students who often gain the most on the tests are those who have the most to gain (i.e., the academically ill-prepared). Thus, if a college or program wished to show the most value added, it should admit the most poorly prepared students (i.e., those with poor high school grades, poor admissions test scores, and poor course preparation), as has been empirically demonstrated by Banta et al. (1987).

In analyzing change in the learning of college populations there is the additional problem of dropouts. That is, it may be that only the students who do, in fact, gain, will still be attending the institution when the final assessment is made. The institution will then appear to be effective, simply because the students who would supply counter-evidence are no longer there.

Another very difficult problem is the attribution of gain to the institution's programs, when it may be due to maturation, the general college environment, or simply to the fact of college attendance. In fact, it appears that students "gain" at about the same rate, wherever they are attending (Baird, in press; Pascarella, 1985). This problem of attribution is especially vexing when the subject of assessment is "general education." It becomes even more complex when the interactions of student characteristics and institutional influences are considered. That is, a student's gain could be influenced by different teaching styles, the student's learning style, the overall quality of instruction in a program, the match of student interest with the major, the overall college curriculum, the student's peers, and the global college environment (Baird, in press).

Practical Issues

There are many problems with putting value added into practice. The first is the delineation of content. This task is difficult enough in a single course, as every professor knows; but difficulties mount when mastery of a major is assessed, and reach their zenith when general education is addressed. How can one define content that will be fair and educationally meaningful to students in different majors? How can faculty agreement be reached? These questions can be answered more easily in small colleges or in institutions with a single focus, than in larger, multipurpose institutions. The more general problem is the unit of analysis, that is, whether one studies educational gain at the level of the course, the major field, the major subdivision (e.g., fine arts, engineering) or the institution.

Another potential problem is the content and difficulty level of the instruments used as the criterion measures of gain. A community college which accepts all applicants who have a high school diploma or a G.E.D. would clearly want to use different criteria of gain than would a selective institution. The community college would use instruments that are relevant to its programs, such as occupational competency assessments or, for its students planning to transfer to four-year colleges, the CLEP General exams. A selective institution would wish to use instruments that are relevant to its programs, which enroll large numbers of students planning graduate degrees. These instruments might consist of tests designed for admission to advanced study, such as the Graduate Record Examinations General Test, or a test of familiarity with abstract conceptual vocabulary and general knowledge expected in traditional liberal education, such as the Concept Mastery Test. Colleges should take the choice of a criterion measure very carefully. In order to be useful, the measure should be matched as closely as possible to the educational programs and goals of the institution.

Care should also be taken to observe the purposes for which tests were originally designed. For example, the SAT and the ACT are designed to assess students' preparation for college. It is very questionable to use these tests as criteria for growth in the college years. As Adelman points out in his essay on general education assessment, the choice of a test is a public statement of the institution's educational objectives. To use the SAT or ACT college admission tests at the end of the freshman or sophomore year could be construed as a statement that a college's objectives for student learning are pre-collegiate. In this same context, the level of difficulty should also be examined carefully, for if the test is too difficult or too easy, the difference in scores for students with different educational experiences will be due more to chance than if the tests were of appropriate difficulty. There are no hard and fast rules for deciding which test to use, but the test should not just be pulled "off the shelf."

Just as difficult as the question of what content to include is the question of how to assess its mastery, that is, whether "objective" multiple choice examinations, essays, or integrative projects best assess such mastery. Obviously, different kinds of assessments are appropriate for different majors and programs. A music performance major should be assessed by performance, understanding of the background of the piece, the quality of the rationale for his or her approach to performance. An English major should demonstrate the capability to write as well as to demonstrate knowledge of literary history and ability to analyze texts from the major literary genres. The problem is that some assessment methods are much more reliable and less subject to the vagaries of human judgment than others. The estimates of the reliability of multiple choice examinations are usually higher than those for essay examinations, which, in turn, are usually higher than those for demonstrations or projects.

A related issue is the degree of specificity of the assessment. An assessment that provides detailed information about each learning objective will be much more useful than a global assessment. However, the effects of some instructional programs (e.g., in writing) probably can be assessed best through global holistic criteria.

A related question is the definition or meaning of a "program." In order to assess the value added by a program, it must have identifiable elements that can be assessed. Many programs have so much flexibility in their requirements that it would be difficult to come to terms with the variety of students' experiences. For example, the degree requirements for an accounting major may demand that students take the same courses, constituting 80 percent of their classwork. Another department on the same campus, say communication, may require courses that constitute only 20 percent of the students' total coursework for a degree. One would be much more certain of the impact of the accounting major than one would be of the communication major.

Another practical problem lies in translating statistical analyses into guides for decisions. If regression analyses are used on general measures, it is sometimes difficult to assign a very large effect to any particular aspect of the college, largely because students' initial status plays such a large role in their final status. An additional proportion of the final status is determined by the students' quality of effort, motivation, etc. (Pace, 1984). The relatively small remaining proportion has to be parceled out among major field instructional efforts, the general curriculum, the peer culture and the overall environment. Studies of this sort often show a complex pattern with small contributions from different aspects of the college experience, and often have rather minimal significance for policy decisions. These problems are magnified when the initial test differs from the later test.

Perhaps the greatest practical problem with assessing gain on a common criterion is that the method allows (and in some cases may encourage) invidious comparisons. For example, if an institution were to study the gains of an entire class by giving the ACT COMP to the students when they were freshmen and again when they were seniors, the average scores would suggest that most students had made gains. The problems begin when the gains of students in different majors or programs are compared. By definition, the gains of some students will be above the average and some will be below. Likewise, the gains for students in some majors or programs will be above or below this average. Thus some programs will be above average and some will be below average. Although these differences may be due to unreliable gain scores, they may be interpreted as differences in the educational effectiveness of the programs. Note that one can reach this conclusion independent of observation of a department's curriculum and instruction in action. Programs or departments that provide quality education can still be considered below average by the value added analysis, and be subject to the attendant administrative penalties.

These invidious comparisons can be made across State systems as well, with institutions that may be providing good educations being judged below average. It is possible that less wealthy institutions, which may be understaffed and ill-equipped, will be "below average" by this definition and deprived of needed resources. Whenever rewards are tied to performing "better than average," the pressure is on to score as high as possible. These pressures can lead an institution to lose sight of its educational purpose. In such a case, practices may evolve that corrupt the assessment process (e.g. "teaching to the test" or admissions policies based on the likelihood that the students will show gains).

An Evaluation of Value Added

The research reviewed by Pascarella (1985) and Nucci and Pascarella (1987) suggests that, as a research tool, the general idea of value added can lead to helpful insights. Very roughly, the research indicates that the largest effects on student growth and change are due to maturation, followed by effects due to attendance at any college, followed by effects due to attendance at a particular college and, lastly, effects due to within-college experiences. Thus, as a research approach, the basic idea of value added, or differential effects, has provided some useful perspectives at the national level.

These same perspectives, however, lead to reservations about the use of value added at the institutional level. It is not clear whether a gain in test scores would be attributable to students' maturation, the experience of attending college, the overall college experience or the particular course or program the students had taken. It seems plausible, however, that if an institution or program has very explicit educational goals, the gain could be attributed to the institution or program. Given the diversity of institutions and programs, comparisons of relative impact would be quite tricky, especially when used to allocate monetary or organizational rewards. Thus, although a value-added assessment strategy may have some utility as a way of examining the educational effectiveness of programs for the purpose of internally generated improvements, it must be done very carefully, keeping the points discussed in this chapter in mind.

As Fincher (1985) has written: "If educators could agree on the assessable outcomes of higher education, take the time and effort to develop suitable forms of measurement and assessment, and restructure instructional efforts in terms of explicit instructional objectives, value-added concepts of education might then be a solution to some educational problems." If executed thoughtfully, value added-assessment has some potential for the improvement of instruction at the program level. It is much less appropriate or useful at the institutional level of analysis. It is, above all, not a panacea, or even a solution to be recommended widely.

References

Astin, A.W. "Why not Try Some New Ways of Measuring Quality?" Educational Record, vol. 63 (1982), pp. 10-15.

Astin, A.W. and Panos, R.J.. The Educational and Vocational Development of College Students. Washington D.C.: American Council on Education, 1969.

Baird, L.L. "The College Environment Revisited: A Review of Research and Theory." In J.C. Smart (ed.), Higher Education: Handbook of Theory and Research, Vol. IV. New York: Agathon Press, in press.

Banta, T.W. et al. "Estimated Student Score Gains on the ACT COMP Exam: Valid Tool for Institutional Assessment?" Review of Higher Education, vol. 27 (1987), pp. 195-217.

Carver, R.C. "Two Dimensions of Tests: Psychometric and Edumetric," American Psychologist, vol. 29 (1974), pp. 512-18.

Fincher, C. "What is Value-Added Education? Research in Higher Education, vol. 22 (1985), pp. 395-398.

Goldstein, H. "Measuring Changes in Educational Attainment Over Time: Problems and Possibilities," Journal of Educational Measurement, vol. 20 (1983), pp. 369-78.

Greenwald, D. and Associates. The McGraw Hill Dictionary of Modern Economics, (3rd ed.). New York: McGraw-Hill, 1983.

Harris, C.W. (ed.). Problems in Measuring Change. Madison: University of Wisconsin Press, 1963.

Kessler, R.C. and Greenberg, D.F. Linear Panel Analysis: Models of Quantitative Change. New York: Academic Press, 1981.

McMillan, J.H. "Enhancing College Students' Critical Thinking: A Review of Studies," Research in Higher Education, vol.26 (1987), pp. 3-30.

Northeast Missouri State University. In Pursuit of Degrees With Integrity: A Value-Added Approach to Undergraduate Assessment. Washington, D.C.: American Association of State Colleges and Universities, 1984.

Nucci, L. and Pascarella, E.T. "The Influence of College on Moral Development," in J. Smart (ed.) Higher Education: Handbook of Theory and Research, Vol. III, New York: Agathon Press, 1987, pp. 271-326.

Nuttall, D.L. "Problems in the Measurement of Change," in D.L. Nuttall. (ed.), Assessing Educational Achievement. Philadelphia: Falmer Press, 1986, pp. 153-167.

Pace, C.R. Measuring the Quality of College Student Experiences. Los Angeles: Higher Education Research Institute, 1984.

Pascarella, E. "College Environmental Influences on Learning and Cognitive Development: A Critical Review and Synthesis," in J. Smart (ed.), Higher Education: Handbook of Theory and Research. New York: Agathon Press, 1985, pp. 1-61.

Pascarella, E.T. "Are Value-Added Analyses Valuable?" Paper Presented at the 1986 ETS Invitational Conference, New York.

Smith, C.K. "The Glitter of Value Added," Research in Higher Education, vol. 25 (1986), pp. 109-112.

Southern Association of Colleges and Schools. Resource Manual on Institutional Effectiveness. Atlanta: author, 1987.

Traub, R.E. and Wolfe, R.G. "Latent Trait Theories and the Assessment of Educational Achievement," Review of Research in Education, vol. 9 (1981), pp. 377-435.

Wood, R. "The Agenda for Educational Measurement," in Nutall, pp. 185-204.

Zoffer, H.J. "Accreditation Bends Before the Winds of Change," Educational Record, vol. 68 (1987), pp. 43-44.

Computer-Based Testing:
Contributions of New Technology

by Jerilee Grandy

In recent decades, there has been a monumental growth in the number and kinds of academic and vocational skills students may choose to acquire in American colleges and universities. Subjects that were once relegated to vocational or business schools are now part of two-year and four-year college curricula. Growth in science and technology has played a major role in the expansion of the college curriculum. Even a secretary must know word processing to have a competitive edge in the job market, and a prospective undergraduate physics major in some universities must choose his or her area of specialization by the sophomore year. With the number of subject areas and major fields growing, there is an increasing need for appropriate measures of knowledge and performance in these fields.

Not only has there been an increase in the areas of learning that educators must assess: there is greater insistence that we also measure higher order thinking skills. The shift in emphasis from fact learning and the development of deductive reasoning skills to "whole-brain" learning has placed an exceptionally great demand on test developers and psychometricians attempting to define and measure learning outcomes.

Paralleling changes in curriculum content and emphasis has been a rapid advancement in computer technology and its application both to instruction and to testing. As a result, educators are no longer limited to paper-and-pencil media, nor are they willing to be limited by them. In most commercially available systems, assessment is designed to facilitate instruction--not as an end in itself. This chapter will therefore explore some of the ways that computers may be revolutionizing both teaching and testing.

The Value of Computer-Based Testing

Any method of testing in which the examinee interacts directly with a computer is called computer-interactive testing. One of the great ironies of computer-interactive testing is that it can return us to an earlier, indeed an ancient, method of assessment, namely, the individualized one-on-one examination. While many people in today's world view computerization as impersonal, computer-interactive testing can be, by this interpretation, wholly personal.

Consider for a moment how an examiner, a classroom teacher, or an individual tutor from earlier times might have determined a student's knowledge of a subject. He might have started with a question of only moderate difficulty, just to get an idea how advanced the student was. The student missed the answer; the question was too hard. So the examiner asked another question and the student answered correctly. The next one was just a little harder, and the student got it wrong again. After a few more carefully chosen questions, the examiner had a fairly clear idea of the student's level of mastery in that subject.

The next student was also given a moderately hard question as a start. She got it correct, so the examiner gave her a harder one yet. She got that one correct as well. The next one, however, was too difficult. So the examiner gave her a slightly easier one and she got it correct. Her level of knowledge was also determined with only eight or ten questions.

Under this system, the examiner could present different questions to different students, so there was no problem with test security. The bright students were neither burdened nor bored by answering dozens of easy questions, and slower students were neither overwhelmed nor discouraged by dozens of questions they could not answer. So long as the examiner was knowledgeable, fair, and reasonably compassionate, the system probably worked quite well. Issues of test validity, reliability, and the nonstandardization of test items were probably not raised. Undoubtedly, performance anxiety, possibly bordering on panic, did occur, especially when entrance to a prestigious university was at stake. But testing could proceed swiftly--until there were hundreds or thousands or tens of thousands of students to be examined. Then mass testing became necessary.

The urgent need for mass testing arose during the First World War when the Army had to select thousands of new recruits for specialized training as efficiently as possible. The Army Alpha was the first large-scale, broad-range examination. Questions were posed in a multiple-choice format so that they could be scored quickly and easily. Since everyone took the same test, the test taker was not subject to the moods or biases of the examiner.

While multiple-choice tests met the needs of the military and of the growing numbers of schools and colleges, educators were becoming increasingly aware of their limitations. Because they were administered to a population with a broad range of abilities, tests like the Army Alpha were dominated by items of average difficulty, but also included some very easy and some very difficult items. This was a fine arrangement for the average person. But it had little reliability at the extreme ends of the ability spectrum. It was not so useful for the evaluation of recruits who were barely literate or those who had doctorates.

Other drawbacks of the format soon emerged. Examinees sometimes made errors gridding the standardized answer sheets by offsetting their responses one row or column, or by marking too lightly or failing to erase completely. Educators discovered that the standardized paper-and-pencil test could be stolen. Once test security was violated, the entire test battery had to be redesigned. Later, the problems of special students began to come to our attention. How can we test the knowledge of someone with a reading disability, when, after all, a test item in any subject has to be read before it can be answered? How can someone with a physical disability, such as blindness, be tested?

Complaints about mass testing multiplied. Can questions in a multiple-choice format really measure important abilities or merely recognition and recall? How can we elicit a creative response to a complex problem and evaluate it using a multiple-choice question? Paper-and-pencil tests were undoubtedly useful for assessing basic skills and factual knowledge. But what about judgment? How do we assess judgment? Or interpersonal skills? Or speaking and listening skills? How do we test whether a beginning teacher can manage a classroom, or who will be an empathic psychotherapist? How can we know which critical care nurse will respond appropriately in a crisis? How can we assess an engineer's ability to troubleshoot a circuit problem, or a programmer's skill in debugging computer programs? Frequently the most critical skills of the most educated people are not easily assessed--and perhaps cannot be assessed--by a paper-and-pencil test. Often the cognitive functions that we would like to measure are complex, and to describe the problem adequately may require many pages of text, or it may not even be possible to present the problem adequately in writing. Sometimes graphics or a motion picture with sound can better communicate the complexities of a problem or its environment.

The heavy reading load on subject tests is evident from the high intercorrelations generally obtained between sub-scores on a battery of diagnostic or achievement tests. It is impossible to measure a student's knowledge of science, geography, or American history (particularly at introductory levels) without the student's reading skill affect the subject score. Tests are simply not reliable enough to allow us to filter out reading ability from the student's knowledge of a specific discipline. Students are at a disadvantage if they are slow readers, if they have reading disabilities, or if their English-language proficiency is limited. They may have mastery over their subject--especially a technical subject or a task requiring spatial or psychomotor skills--but they may not be able to perform well on a written test.

In very recent years, progress in the use of computers for teaching and assessment has enabled educators to grow beyond many of the limitations of traditional testing. A simple example is the replacement of the answer sheet with an electronic recording system that provides immediate feedback and scoring. A single microcomputer is coupled with keypads for each student, and produces individualized print-outs of responses and scores. The system can obviously be used for instruction as well as assessment. This is but one example of a low-cost innovation based on very simple technology. The remainder of this chapter will review three ways that computer technology has addressed the more knotty problems inherent in paper-and-pencil testing. One way has been to reduce the length of multiple-choice tests by using computerized, adaptive testing. A second has been to introduce new media, such as graphics and video segments, to replace written text. Finally, the use of expert systems to score constructed-response test questions is currently in its research and development phase.

Computerized Adaptive Testing

For a college or university to assess student achievement with adequate reliability in many fields, paper-and-pencil tests must be long and require much of the examinee's and examiner's time. (For a discussion of the relationship of test length to reliability see a standard textbook on measurement such as Anastasi, 1988.) Adaptive testing applies the statistical methods of Item Response Theory (IRT) to tailor the difficulty of a test to the skills of individual examinees (See Wainer, 1983, and Lord, 1980). As a result, it provides very efficient measurement across a broad range of levels of skills with a test about half as long as its paper-and-pencil equivalent.

The simplest form of an adaptive test works in the following way:

Step 1. The computer chooses at random one of the middle-difficulty-level questions and presents it to the examinee.

Step 2. Depending on whether the examinee answers correctly or not, the computer randomly selects another item from either the easiest or the most difficult questions.

Step 3. The computer continues to monitor the examinee's responses and to present questions of appropriate difficulty until it "zeroes in" on the student's skill level.

A test presented in this format can have a great many advantages over the traditional paper-and-pencil test:

o For program evaluation requiring a pretest/posttest design, alternate forms of a traditional test are generally necessary so that students do not receive the same test twice. The same adaptive test can be administered to the same students many times because the item-selection algorithm will (with high probability) select different items each time the student takes the test. If the student's knowledge increases between pretesting and posttesting, the program will branch the student to more difficult items so that, in essence, the student receives an entirely different posttest.

o Testing time may be considerably shorter than with a paper-and-pencil test because the computer program determines which questions will be given to the individual, and the program eliminates questions that fall outside the examinee's ability range. In that way it determines the examinee's skill level with a minimum number of questions. The examinee's score is determined not by how many questions are answered correctly but by which questions are answered correctly.

o Students report that they prefer adaptive testing because it is shorter, because they are not bored by having to answer questions that are too easy, and because they are not anxious or discouraged by attempting large numbers of items that are too difficult.

o It provides fine discrimination over a wide range of ability levels. Conventional tests have high measurement precision near the average test score, but have low measurement precision for scores at the high and low ends of the scale. Precision over a wider range of scores can be assured only by lengthening the test with a greater number of very easy and very difficult items. Adaptive tests, on the other hand, maintain high precision or accuracy at all ability levels. By setting the termination criterion at a specified value, all examinees may be measured to the same level of precision.

o Like other kinds of computer-interactive testing, adaptive testing permits immediate reporting of test results.

o Test security is improved with adaptive tests, even over other forms of computerized tests, because each examinee receives an individualized test. It would be difficult to memorize the items from all of the possible tests. Two examinees sitting next to one another are most unlikely to see the same item at the same time. There are no answer books or answer sheets to be stolen. Encryption can be used to protect item banks and individual student data.

A number of major organizations have been involved in developing accessible computer adaptive tests. The military has sponsored the most far-reaching and complex projects, including the computerized adaptive administration of the Armed Services Vocational Aptitude Battery (ASVAB). The Educational Testing Service and the College Board have produced the College Board Computerized Placement Tests in Reading Comprehension, Sentence Skills, Arithmetic, and Algebra (College Entrance Examination Board, 1985). These untimed computerized tests automatically produce scores and a variety of score reports and summaries. They can discriminate better and more reliably than paper-and-pencil tests but take only about half as much time to administer because they contain only 12 to 17 items each. And the Psychological Corporation has developed an adaptive version of the Differential Aptitude Test (DAT) for administration on Apple II computers (Psychological Corporation, 1986). The DAT is a battery of tests suitable for high school students and adults. It provides scores in verbal and numerical abilities, abstract reasoning, clerical abilities, mechanical reasoning, and other aptitude areas.

Despite these successful adaptive tests currently in use, there are a number of problems with the method. Even test developers and psychometricians experienced with adaptive testing and item response theory cannot always produce a satisfactory test.

Problems with Computer Adaptive Testing

Development of an adaptive test requires great care in the calibration of a large pool of items, possibly several hundred. These items are administered to a large number of examinees from the target population, and using item response theory, so-called "response vectors" are obtained for each examinee. Using the data from the response vectors, calibration programs estimate the parameters of the item response curves. Once calibrated, these items are retained in an item bank. New items must constantly be added to the item banks. Furthermore, it is necessary to use a special item analysis technique to study difficulty parameters, discrimination parameters, and guessing parameters in order to refine or discard new items.

One problem with adaptive tests is that they are more vulnerable to "context effects." This means that a student's performance on one item can be affected by that item's relationship to other items in the test. If all examinees are presented with the same items, the context effects are assumed to be the same for everyone. But with different students receiving different items, the context will not affect everyone similarly.

Another type of context effect arises if a particular theme or subject appears repeatedly. In a traditional paper-and-pencil test, test developers now avoid unbalanced content, especially on socially sensitive subjects. For example, if a reading passage uses a male first name for a person in a technical job, an item writer will take care to use a female name in another item having a person with a similar role. Likewise, if one sentence completion item refers to a traditionally female type of recreation, such as dance, another item may refer to a traditionally male sport, such as basketball. In adaptive testing, not all examinees are presented with the same items. One person may, by chance, encounter all mostly "male" items, while another encounters mostly "female" items. A similar kind of imbalance would occur if one examinee got two successive items in which the main character had an Hispanic name, or several vocabulary items drawn from literature and none from science.

Another problem is that if an item is flawed in some way, its detrimental impact is greater on an adaptive test than on a traditional paper-and-pencil test of greater length. This variance occurs because the shorter test lacks the redundancy inherent in a longer, conventional test. If an item fails to perform as expected, its detrimental impact on validity may be twice as great as would be the case on a traditional test of the same accuracy but greater length.

Other problems may arise due to the order of presentation of items. A number of studies have found differences in the apparent difficulties of items depending on their location in the test. Studies have also shown that when items are arranged in order of decreasing difficulty, instead of the typical order of increasing difficulty, the overall effect is to increase the difficulty of the test, probably by increasing anxiety and frustration. In adaptive testing, the lower ability examinees are first presented with an item of medium difficulty, which for them is of high difficulty. The computer then presents them with successively easier items. The net effect for those examinees may be a test containing items of decreasing difficulty.

Psychometricians are also concerned about the factor structure (what the test actually measures) of an adaptive test. A study by Green (1987) showed that when some paragraph comprehension items were modified for computer presentation, they behaved more like word knowledge items. The factor structure may also change over time as the student's knowledge increases. The use of change scores to measure achievement over time assumes that the pretest and posttest are measuring the same skills or knowledge, that is, the factor underlying changes in performance is invariant over time. Gialluca and Weiss (1981) found that this was not always true, however. They found that the factor structure of achievement in a biology course was not the same before instruction as it was several weeks after instruction. In a mathematics course, however, the factor structure remained the same over a ten-week period.

"Testlets"

While there are still problems in the application of computerized adaptive testing, these problems do not preclude its further development. Wainer and Kiely (1987) have been applying a multi-stage fixed branching adaptive test model that substitutes multi-item "testlets," or very short tests, for single items. They define a testlet as a "group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow." Because each item is embedded in a predeveloped testlet, it essentially carries its own context with it.

The fixed branching testlet avoids, or at least minimizes, the problems inherent in the variable branching adaptive test models currently under development. It provides only a limited number of paths for examinees to follow, and therefore, if test developers construct each path carefully, they can avoid most of the problems identified earlier.

While the design of testlets has not yet become a standardized procedure, a section of the Architect Registration Examination on seismic/lateral forces has been redesigned by ETS for the National Council of Architectural Registration Boards (NCARB) as a computerized mastery test (CMT) utilizing testlets. It has been successfully pilot tested, and will be field tested on a large scale in the near future. The CMT is not strictly an adaptive test. Rather, it allows the examinee to answer only as many questions as are needed to determine whether he/she passes the test. All students, however, take the same test items, i.e., they do not branch to harder or easier items as they do in an adaptive test. If the examinee's performance on the first 20 items is far from the cutting score (either higher or lower), there is no need for that examinee to answer more questions. A pass/fail decision can be made with little probability of error. If the examinee's performance lies near the cutting score, the computer presents another set of questions in order to obtain a more precise estimate of the examinee's knowledge. This process is repeated until a reliable pass/fail decision can be made. An examinee performing very near the cutting score will take the entire 60-item test.

Use of New Media

Reading has always been the standard medium for conveying information, especially in academic learning. What the new technologies are offering (with considerable resistance from some educators) are alternatives to reading. Computer graphic displays have many advantages over descriptions and drawings, the major one being the ability to change and therefore to represent a process. A question aimed at assessing whether a student understands the role of messenger RNA in the synthesis of a polypeptide can be communicated more clearly in a moving graphic display than in a paragraph accompanied by a diagram.

Even more realism can be achieved by audiovisual techniques, especially when human interaction is presented. It may be important for the examinee to be able to interpret the subtleties of body language and voice intonation, or to comprehend a foreign language or dialect. One medium for this kind of presentation is interactive videodisk. The technology for interactive videodisk has been in use for instruction, especially in continuing education, for the past decade. Videodisks have a considerable advantage over videotapes in that they allow for branching to different video segments rather than requiring the viewer to watch a tape from beginning to end in a linear fashion. This capability allows the viewer to interact with the computer program.

The program may first present an instructional video segment, such as a teacher managing a classroom, a doctor treating a patient, or a police officer arresting a suspect. The program may then present a question or series of instructions related to the video segment and wait while the viewer responds. Depending on the viewer's response, the program can determine whether the viewer has understood the previous video segment, and, in accordance with built-in decision rules, it can branch to another video segment that is more advanced (if the viewer has answered the question correctly) or it may branch to a segment that elaborates further on the topic that the viewer misunderstood. Many other kinds of logical sequences are possible.

A combination of audio/video instruction and self-assessment has been applied in a wide variety of educational areas, but it is possible to use the same technology simply for assessment, excluding the instruction. Video technology makes it possible to incorporate realistic sound and motion into assessment. A computer program can show a video segment followed by a question or instruction. The viewer can then respond. The program can then evaluate the response and either branch to another video segment, print a message back to the viewer, produce a score for the response, or do essentially anything that the designer wishes it to do. Using a series of video segments followed by questions and answers, the program can lead the viewer through an entire examination, produce a score report, and even explain to the viewer where his or her strengths and weaknesses lie. The methods of adaptive testing, discussed earlier, can be built into the testing and scoring system.

Alternatives to Multiple-Choice

Perhaps the greatest challenge to psychometricians, cognitive psychologists, and computer scientists in the area of testing is the development of an "intelligent" computer program that could test a student the way a perfectly knowledgeable, fair, consistent, understanding, and objective examiner would. The computer would ask a question, wait for a response, and then judge the accuracy and quality of the response on the basis of objective rules. The student's response would, if possible, be a "free" or "constructed" response, meaning that it would not be limited to multiple choice, but would be created by the student. Ideally, the computer would also function as an "expert system" and supply explanations and justifications to the user and perform other complex functions.

At the present time, considerable effort is going into the development of expert systems, also called knowledge-based systems. Emerging from the field of artificial intelligence (the study of principles of computational intelligence manifested in man-made systems), research on Intelligent Tutoring Systems (ITS) is attempting to translate the "art" of teaching into a "science" of teaching. More specifically, it is attempting to translate the instructional process into a set of systematic algorithms. While many educators would deny that such a reduction is possible, others point out that the "art" of medical diagnosis (Shortliffe, 1976) and the "art" of designing synthesis paths for complex chemical molecules (Wipke et al., 1977) have been shown to be reducible to systematic algorithms. It should be possible, they argue, to codify the instructional process as well.

Anderson and his colleagues at Carnegie Mellon University have successfully designed a tutor, based on a learning theory known as ACT, that teaches LISP (a computer language used to design expert systems). The tutor provides instruction in the context of problem solving, has the student generate as much of each solution as possible, provides immediate feedback on errors, and finally, represents the structure of the problem for the student. Anderson and Reiser (1985) showed that students using the tutor performed better than students in a standard classroom and nearly as well as students with personal tutors. For more technical information on the LISP tutor, see Anderson, 1983; Anderson and Skwarecki, 1986; and Reiser, Anderson, and Farrell, 1985.

The key to the tutoring program's success is its ability to fit each student response into a model of correct and incorrect methods for solving a problem. The tutor must be able to analyze each portion of the student's solution in order to diagnose errors and to provide guidance. This process of understanding the student's behavior as it is generated is called "model-tracing" (Reiser, Anderson, and Farrell, 1985). By following a student's path through a problem, the tutor always has a model of the student's intentions. According to Johnson and Soloway (1984), inferring intentions is necessary for responding appropriately to students' misconceptions about problem solving. Soloway and his colleagues have developed a computer-based expert system called PROUST that can accurately detect and diagnose student errors in PASCAL programs (Johnson and Soloway, 1983). They are presently exploring other applications of the same type of system, and have successfully applied it to the scoring of statistics problems.

Technology employing expert systems is still in its infancy. Cognitive psychologists still do not fully understand the problem solving process or the ways that people make errors. Many argue that it is impossible ever to individualize instruction fully. The work of Cronbach and Snow (1977) suggests that Aptitude Treatment Interactions cannot be generalized sufficiently to be useful as guides to instruction.

In addition to the limitations imposed by our lack of understanding of the learning process, psychometricians have not yet developed a measurement theory that will encompass constructed-response formats. Given the current state of technology and theory, a test using an expert system may be most valuable for a qualitative rather than quantitative diagnosis of students' strengths and weaknesses. If a vast majority of the students taking a chemistry test, for example, fail to answer the questions on molality correctly, there would be strong evidence that the chemistry program is faulty in its teaching of molality. A well-designed expert system might point out more specifically that students are confusing molal and molar solutions and that they are consistently making errors in problems that rest on this distinction. More detailed information on intelligent tutoring systems can be found in two recent books, one by Wenger (1987) and the other edited by Lawler and Yazdani (1987).

Computer-Based Item and Test Construction

Computer-based testing applications are not restricted to the administration and scoring of tests. A test developer can use computers to create and store items and to construct tests.

Item Banks

The simplest use of a computer for test construction is to employ it as an item bank. Then, when a test has to be constructed, items are chosen according to a predetermined rule or, if they are presented interactively, the order of item presentation depends upon the examinee's response to the previous item.

A second way to make use of computers for test construction is to store rules for item generation, rather than the items themselves, in the computer. Fremer and Anastasio (1969) were among the first to use this algorithmic approach to test item generation. Others, e.g. Roid and Haladyna (1982) have since developed algorithms for generating items for every kind of subject matter.

In mathematics, an algorithm for computing sets of simultaneous equations, for example, could produce infinitely many items of the same form but containing different numerical values. Any item selected at random would test the student's ability to solve simultaneous equations. Algorithms for creating items from reading passages have also been developed based on specific kinds of linguistic transformations.

For further information on the use of computers for item and test construction, see Millman (1980 and 1984). Computer generation of spatial items is currently being researched by Bejar and Yocom (1986).

Practical Considerations

The development of computer-based tests is a gradual process whereby a small, manageable prototype is designed, field-tested, revised, and implemented. Most often, a second test is designed along similar lines in a shorter period of time and at less expense. Once in use, a test is seldom regarded as "finished." This is especially true of expert systems which, by their very nature, are "learning" and improving their ability to "understand" the errors made by examinees. The decision to computerize testing, therefore, does not have to entail a major commitment to a new way of testing. It may be accomplished in small increments enabling the designers to re-evaluate and improve program segments regularly and to update hardware as technical improvements become available.

The development of any test requires specialized expertise in the content domain of the test, item writing, and psychometrics. Computer-based test development demands additional expertise, much of which is available in a major college or university. Designing an adaptive test requires knowledge of item-response theory and the use of special computer programs. It may require additional statistical expertise. The architecture test for NCARB, for example, required the development of special extensions of Bayesian statistics.

A computer-based test can be written in a standard programming language such as FORTRAN. Special languages such as LISP and PROLOG have also been developed because FORTRAN, BASIC, and other traditional programming languages are cumbersome and impractical for expert systems design. Programming skills are not necessary, however, for the design of computer-interactive tests. With currently available test construction software, or authoring systems, a person with no training in programming can select the topics and objectives for the tests to be created, review the specifications, select specific items to be used in the test, sequence the objectives and items, and create an operational test to be administered by computer. By following menus or working from promptlines, the user can create text and graphics, edit video, and animate areas of the screen. Often the software includes all the necessary tools for test registration, scheduling, management, administration, scoring, reporting, and providing specific curriculum prescriptions. These applications are discussed by Olsen et al. (1984) and by Slawson et al. (1986). The great advantages of authoring systems are that they require no programming skills and are less time consuming and costly to use than a general programming language. On the other hand, because they are easy to use, they are less flexible. Whether a computer-interactive test is designed with authoring courseware or in a general programming language depends, therefore, on the programming expertise of the designer and on the flexibility that the program demands.

There are countless authoring systems on the market, and the popularity of each seems to rise and fall rapidly as it is superceded by a "better" system. CDS/GENESIS is a currently popular authoring system available from Interactive Technologies Corporation of San Diego; QUEST is a well-known authoring system from Allen Communication of Salt Lake City. General catalogs of courseware are also available. IBM and Sony both publish catalogs of videodisk courseware and authoring tools. Applied Video Technology Inc. of St. Louis has also published a directory entitled Interactive Video that includes users, vendors, and producers concerned with interactive videodisk.

The time and cost required to build and operate a computer-interactive test depends on a great many variables. Twenty years ago, an expert system could require from 20 to 30 person-years to develop. Today, that figure has been reduced to less than 5 person-years. To design an expert system, a major effort must be made to identify the kinds of errors an examinee might commit and to translate error patterns into useful diagnostic information by way of production rules. The usefulness and validity of the test depends on the accuracy and completeness of those production rules which, in turn, depend on the time and effort that went into their design.

The development of an adaptive test using multiple-choice items may take two person-years or more if original test items must be written and calibrated. Development may take considerably less time if a paper-and-pencil test already exists and the items have already been administered to several hundred students. Using the existing test data, the items can be calibrated. Only the program to administer the test has to be designed.

Hardware requirements for administering a simple multiple-choice adaptive test are minimal. Any standard personal computer with a black and white monitor and floppy disks can be used. If the nature of the test is such that color graphics or high-resolution graphics are desired, a more expensive monitor will be necessary. Generally a hard disk is desirable.

Some tests, especially those employing expert systems, may lend themselves to the use of a voice synthesizer, music synthesizer, or other specialized equipment. They may require more than the standard screen and keyboard. To troubleshoot an electric circuit problem, an examinee may have to trace the circuit using a mouse, light pen, or touch screen. Additional hardware obviously increases costs.

Rarely is there a need for more storage area than a hard disk supplies. Skeptics of computer-based testing often doubt that a test with graphics, a large knowledge base, and many production rules can be run on a personal computer. If there is not sufficient space on a hard disk, the system can use a CD-ROM (Compact Disk Read Only Memory). This small disk can store the equivalent of more than 1,500 floppy diskettes. To illustrate how much information that contains, consider that an entire 20-volume encyclopedia, the Academic American Encyclopedia, which contains over 30,000 articles, is contained on a single disk. A low-cost CD-ROM player is currently being introduced by Atari for under $600.

The development of interactive videodisk can be difficult and quite expensive, especially if original filming must be done. The cost of scripting and filming may exceed $100,000. One reason for the high cost is that students are accustomed to professional quality film productions. An amateur production of a social interaction, for example, may not be taken seriously by examinees. If professional actors must be hired, costs can be high. Production can sometimes be done easily, however, if relevant video segments are already available, such as those used in a course. A college may also consider calling upon the resident expertise of its communications and drama departments, where students may be able to create a highly effective production.

Videodisk hardware is relatively expensive (compared with computer graphics, for example) and the disks themselves can be recorded only once. The cost of hardware, however, is trivial compared with the cost of production, and because of competition, the hardware costs are steadily decreasing. Sony, for example, now produces a videodisc player for under $1,000. Applied Interactive Technologies of Jackson, Miss. advertises an interactive video machine for $1,500. It includes a laserdisc player, keyboard, and dedicated microprocessor, and can be attached to a standard television.

There are a countless number of instructional programs on the market, and although they have assessment components that guide the program in tailoring instructional material to the needs of the learner, the primary emphasis of these programs is on instruction rather than assessment. Some universities are beginning to develop sophisticated tutorials employing expert systems, but many educators believe that it is wasteful to use such a powerful teaching tool merely for testing. In the near future, we will probably see modular instructional packages with pretest and posttest assessments for each module. As the student progresses through the hierarchy of learning modules, he or she will have a measure of progress at each stage. That measure will be useful not only to the student, but it will provide useful feedback to the instructor, and it will serve as input to an institutional assessment system.

Until more research and development is done on computer-interactive testing, the technology will not replace paper-and-pencil testing for some time. But its applications are growing daily, and it is certainly appropriate to consider a computer-based test as the first building block in a higher education assessment program.

References

Anderson, J.R. The Architecture of Cognition. Cambridge, MA: Harvard University Press, 1983.

Anderson, J.R. and Reiser, B.J. "The LISP Tutor," Byte, vol.10 (1985), pp. 159-179.

Anderson, J.R. and Skwarecki, E. "The Automated Tutoring of Introductory Computer Programming," Communications of the ACM, vol. 29 (1986), pp. 842-849.

Bejar, I.I. and Yocom, P. A Generative Approach to the Development of Hidden-Figure Items. Final report to the Office of Naval Research. Princeton, NJ: Educational Testing Service, 1986.

College Entrance Examination Board. Computerized Placement Tests: a Revolution in Testing Instruments. New York, NY: College Entrance Examination Board, 1985.

Cronbach, L.J. and Snow, R.E. Aptitudes and Instructional Methods. New York: Irvington, 1977.

Fremer, J.J. and Anastasio, E.J. "Computer-Assisted Item Writing: I: Spelling Items," Journal of Educational Measurement, vol. 6 (1969), pp. 69-74.

Gialluca, K.A. and Weiss, D.J. "Dimensionality of Measured Achievement Over Time," Research Report 81-5. Minneapolis: Computerized Adaptive Testing Laboratory, University of Minnesota, 1981.

Green, B.F. "Construct Validity of Computer-Based Tests," in H. Wainer and H. Braun (eds.), Test Validity. Hillsdale, NJ: Lawrence Erlbaum Associates, 1987, pp. 77-86.

Johnson, W.L. and Soloway, E. "PROUST: Knowledge-Based Program Understanding." Proceedings of the 7th International Conference on Software Engineering, IEEE: Orlando, FL, 1984a, pp. 369-380.

Johnson, W.L. and Soloway, E. "Intention-based Diagnosis of Programming Errors." Paper presented at the National Conference on Artificial Intelligence, Austin, TX, 1984b.

Lawler, R.W. and Yazdani, M. (eds.). Artificial Intelligence and Education. vol. 1. Norwood, NJ: Ablex Publishing, 1987.

Lord, F.M. Applications of Item Response Theory to Practical Testing Problems. New York: Lawrence Erlbaum Associates, 1980.

Millman, J. "Computer-Based Item Generation," in R. Berk (ed.), Criterion Referenced Measurement: the State of the Art. Baltimore: The Johns Hopkins University Press, 1980, pp. 32-43.

Millman, J. "Individualizing Test Construction and Administration by Computer," in R. Berk (ed.), A Guide to Criterion-Referenced Test Construction. Baltimore: The Johns Hopkins University Press, 1984, pp. 78-96.

Olsen, J.B., et al. "The Development and Pilot Testing of a Comprehensive Assessment System." Proposal submitted to the California State Department of Education. Provo, UT: WICAT Education Institute, 1984.

Psychological Corporation. Computerized Adaptive Differential Aptitude Test. San Antonio, TX: Author, 1986.

Reiser, B.J., Anderson, J.R., and Farrell, R.G. "Dynamic Student Modelling in an Intelligent Tutor for LISP Programming." Paper presented at the International Joint Conference on Artificial Intelligence, Los Angeles, 1985.

Roid, G. and Haladyna, T. A Technology for Test-Item Writing. New York, NY: Academic Press, 1982.

Shortliffe, E.H. Computer-Based Medical Consultations: MYCIN. New York: American Elsevier, 1976.

Slawson, D.A., et al. "Waterford Test Creation Package." Provo, UT: Waterford Testing Center, 1986.

Wainer, H. "On Item Response Theory and Computerized Adaptive Tests," The Journal of College Admissions, vol. 28, no.4 (1983), pp. 9-16.

Wainer, H. and Kiely, G.L. "Item Clusters and Computerized Adaptive Testing: A Case for Testlets," Journal of Educational Measurement, vol. 24, in press.

Wenger, E. Artificial Intelligence and Tutoring Systems: Computational and Cognitive Approaches to the Communication of Knowledge. Los Altos, CA: Morgan Kaufmann Publishers, Inc., 1987.

Wipke, W.T., et al. "SECS-Simulation and Evaluation of Chemical Synthesis: Strategy and Planning, in W.T. Wipke and W.J. House (eds.), Computer-Assisted Organic Synthesis. American Chemical Society, 1977, pp. 97-127.

States of Art in the Science of Writing and Other Performance Assessments

by Stephen Dunbar

The purpose of this essay is to outline concerns that pertain to the reliability and validity of writing samples as direct measures of an individual's writing skill. Writing is considered here as a specific example of a performance assessment that yields a product to be evaluated. As a measure of performance, the writing sample is not unlike the recital in music or the field experiment in ecology in that it involves highly complex processes, the product of which is judged by experts in the area. While measurement specialists have similar concerns with respect to any performance assessment, these concerns are more clearly illustrated in the case of writing.

Several contributors to this volume have described basic principles in behavioral measurement, particularly those captured by the terms reliability and validity. The obvious scale of choice for the druggist is the one that gives near identical readings when the same cold capsule is placed on it repeatedly and, moreover, the one that gives close to the correct reading each time. Scales for measurements of writing skill are evaluated by the same criteria. When the scale is one that must be based on an extended sample of performance that yields a single product, as in the case of writing samples, simulation exercises, laboratory experiments, and the like, procedures that specialists typically use to reduce chance errors of measurement and thereby obtain reliable indicators no longer pertain.

In performance assessments, chance errors of measurement can abound because of the time-consuming nature of data collection, the complexity of the scoring task, and the complexity of the behavior assessed. Systematic errors that affect validity are influenced not only by the type of task administered but also by the criteria used in rating the products. Such difficulties render the assessment of writing performance a delicate endeavor. Careful articulation of some of the pitfalls can help faculty committees and administrators in designing a writing assessment that meets local needs while at the same time yields measurements that are trustworthy.

This essay will attempt such an articulation in the context of measuring performance on writing tasks, but with an eye toward implications regarding general problems encountered using production-based measures. Performance tests have long been recognized as important adjuncts to objective paper-and- pencil tests (cf. Ryans & Frederiksen, 1951). A resurgence of interest in them in private industry, the armed services, and now higher education, motivates the present need to broaden understanding of conditions that affect the fidelity of measures obtained from extended observations of behavior. A focus in this essay on direct measures of writing performance is intended to provide the detailed presentation of reliability and validity issues that is necessary for making sense of any performance assessment from the point of view of measurement concerns. Sources of variability in individual scores are at the heart of this matter.

Sources of Chance Variation in the Writing Sample

A sensible discussion of some of the problems posed by direct writing assessments requires a brief, though somewhat stylized, description of a typical assessment because it clarifies elements of the data that influence reliability. Imagine that the engineering college of a large university wants to evaluate the writing skills of its undergraduate majors and devises a series of "assignments" representing the range of writing tasks--office memos, business letters, technical reports, operations manuals--that engineers employed in the private sector encounter as a normal part of their jobs. Specific writing assignments or prompts (the analogues of test items in direct writing assessment) are assembled and administered under controlled conditions and responses are collected and scored by faculty raters, subject to standards of performance outlined in protocols for scoring. Results of the assessment are used to make recommendations for remedial work in a technical writing laboratory being established by the engineering college.

Performance Assessments

Before going further with the example, the reader should note the existence of analogous types of "assessments" in other higher education programs. In music composition or performance, for example, majors are typically required to demonstrate their acquired skill by means of a product that is judged by experts in the discipline. That product may be an original score or a recital, but the properties of the assessment have much in common with those discussed here for the case of writing. In architecture or photography, majors may complete a senior project or display their work at a public exhibit. In teacher education, degree candidates may participate in simulations of the classroom teaching experience in addition to fulfilling practice teaching requirements. These, too, are instances of performance assessment that in some disciplines have evolved into quite elaborate forms of review and evaluation.

In the case of writing, the performance of the engineering students is to be evaluated in several modes of discourse, presumably by one or more faculty raters. The reliability of the individual score, defined as the consistency of that score over repeated observations, is influenced by the consistency of an individual's performance on the specific task and by the consistency of the faculty in providing ratings of that performance. In the example, variability of the observed scores over replications would be attributed to variability in student performance and variability in rater performance. To the extent that student experience in writing, say, operations manuals as opposed to office memos varies, it is conceivable that discourse type itself could influence reliability; however, such variation is usually discussed as it affects the validity of score interpretations (cf. Breland, 1983; Quellmalz, 1986). In any event, such sources of variation are used to explain differences between the observed score and what might be called the individual's true writing ability, differences that the test theorist calls chance errors of measurement. Understanding how to interpret results of a writing sample is in large measure understanding sources of measurement error that can be attributed to chance.

The above characterization should be sufficient to illustrate essential differences between the performance measure and the conventional objective test. Individual variations in performance affect the reliability of both types of instruments. However, the scoring procedures are vastly different; any comparison juxtaposes an optical scanner with a fallible human rater, and a potentially simple response to an objective test item with a product from an extended, highly complex form of behavior. Of course, any test developer knows that objective test items can involve so-called higher-order skills and are thus reflections of complex behavior, but specialists usually acknowledge the distinctive features of production-based measures of performance as well (Lindquist, 1951).

Conundrums of Reliability

One effect of this sharp contrast between objective and performance-based measures has been to concentrate on the consistency of the raters in establishing the reliability of the performance assessment, and for direct writing assessment the development of approaches to scoring essays that produce substantial agreement between judges. Presumably, if scoring procedures can be developed that approximate the consistency of the optical scanner in scoring objective test items, then the additional chance errors associated with the direct writing assessment can be tamed. As discussed later in this essay, concern over the consistency of raters has often altered the focus of investigations of the reliability of the writing score. Preoccupied by raters, investigators have ignored, for example, the consistency of the writer as a source of variation in writing assessments. As a matter of fact, the data needed to estimate the consistency of the writer are not even collected in many direct writing assessments.

A comprehensive report of the results of a recent study of the reliability and validity of direct writing assessments at the undergraduate level is provided by Breland, Camp, Jones, Morris, and Rock (1987). Although the main purpose of the Breland et al. study was to determine the most effective methods for measuring the kinds of writing skills needed for success in college composition courses, the results of this study offer a suitable illustration of the methods for estimating the reliability of writing samples. The 350 participants wrote two essays over the course of a semester in each of three modes of discourse. The two essays were evaluated by three readers in terms of overall quality. The multiple responses and multiple readings design used in this study allowed for estimates of variability in individual performance due to: (1) different essay topics within the same mode of discourse; (2) between-mode differences in performance; and (3) between-reader differences. Each was used in a detailed analysis of the reliability of individual scores. Further description of this approach will provide some flavor for the technical complexity involved in establishing the reliability of any direct assessment of writing skill, and more generally, of any production-based measure.

In the Breland et al. study, descriptions of the reliability of the essay scores were approached in several ways. When more than one reader scores each essay, it is possible to compute a simple correlation between the ratings given by pairs of readers. These correlations are commonly used as estimates of reliability in writing assessments, but tend to give an optimistic picture of score stability because the only source of variation in the responses that influences them is that portion due to inconsistency of the readers. Theirs is a focus on what Millman calls rater reliability. In cases where students write several essays, preferably in the same mode of discourse, a correlation can also be computed between different responses as evaluated by the same rater. Such correlations were reported by Breland et al., who noted that these statistics are sensitive to sources of chance error due to individual performance or topic, but ignore variability due to raters. Hieronymus and Hoover (1987) refer to such correlations as estimates of score reliability.

Analysis of Variance and Reliability Estimates

To obtain reliability estimates that reflect all possible sources of chance variation in individual scores of writing samples, it is advisable to use a more complex approach involving an analysis of variance (ANOVA) design. Such a design allows for estimates of components of individual score variation in such a way that ratios of variation due to a particular source (such as readers) to total variation can be determined and reported as reliability estimates in the form of intraclass correlations (Ebel, 1951). The technical details of this approach are of no great concern for present purposes. The example described below should be understood as an illustration of what the approach offers above and beyond the separate evaluations of rater reliability and score reliability that are more commonly presented as reliability evidence for direct writing assessments.

Table 1, adapted from results of the Breland et al. (1987) study, illustrates some typical properties of reliability estimates for writing samples obtained in the various ways discussed above. The four blocks in the table, labelled A through D, provide reliability estimates obtained by correlational and ANOVA approaches. Each coefficient can be interpreted as a type of correlation, so that the closer the statistics in the table are to 1.0, the more reliable is the overall assessment considered in that manner. Except for the blocks in the table containing the ANOVA results, columns representing more than one reader were determined by an extrapolation method--the Spearman-Brown prophecy formula--that estimates what the reliability coefficient would have been had the writing score been based on multiple readings, using parallel scoring criteria. The main point of the Breland (et al.) results concerns the general magnitudes of correlational estimates, particularly those that attend only to reader variation, as opposed to ANOVA-based estimates. It is clear from the table that

> (1) reading reliability estimates (block A) tended to be greater than score reliability estimates (block B), which directly assessed consistency of performance within the same discourse mode,
>
> (2) the ANOVA-based figures, which reflect all relevant sources of chance error, tended to give the most conservative reliability estimates, and
>
> (3) the estimates based on extrapolation to multiple readings (blocks A and B in the table) tended to be higher than those that were determined from the observed data and design (block C).

Each of these observations is supported by results from other studies of rater versus score reliability in direct writing assessment (cf. Hieronymus & Hoover, 1987; Breland, 1983). When performance was pooled across all discourse modes included in the ANOVA design, so that scores were based on a total of six writing samples, the higher reliability estimates of block D were obtained.

Although the technical details of the ANOVA-based reliability estimates are beyond the scope of this essay, the reasons for preferring them to the correlational estimates usually reported are important. The ANOVA-based estimates of reliability in a performance assessment are usually preferred by

Table 1. Illustration of Types of Reliability Estimates for Writing Samples

Discourse Type	Reliability Type	Number of Readers		
		1	2	3
Narrative	A	.52	.68	.72
Expository		.61	.75	.82
Persuasive		.63	.77	.83
Narrative	B	.37	.54	.63
Expository		.40	.57	.67
Persuasive		.47	.64	.73
Narrative	C	.36	.47	.53
Expository		.40	.49	.54
Persuasive		.46	.57	.62
Pooled	D	.66	.75	.84

Note: A = average correlations between readers, or reader reliability
B = average correlations between essays, or performance reliability
C = ANOVA-based intraclass correlations, reflecting effects of topics within mode, readers, and the interaction
D = all modes of discourse combined

Adapted from Breland, Camp, Jones, Morris, and Rock (1987) measurement specialists for complex studies of reliability because they explicitly identify aspects of the data that contribute to untrustworthy measurements. This is important in practical settings where students write on one of several topics. It is also important when faculty are new to the experience of scoring essays with respect to specific criteria that are not of their own choosing and that they may not fully understand. If essay scores are influenced as much by which topic the writer selected or who evaluated the final product as they are by the skill of the writer, then the scores cannot be interpreted as indicators of writing ability alone.

Besides providing more realistic estimates of chance errors in writing samples, the ANOVA approach lends itself to characterizations of hypothetical assessments that might be designed by a particular institution. Some practical examples may help illustrate the point. Given the estimates of components of variance in the Breland et al. study, it was possible to describe what the overall reliability of the assessment would have been if fewer readers, topics, or modes of discourse had been included. Such estimates, analogous to what statisticians do when recommending sample sizes for detecting experimental effects of a given magnitude, are useful in situations where a pilot study of a particular design for measuring writing skill is performed, but where the feasibility of large-scale implementation is uncertain.

The magnitude of the reliability estimates in Table 1 should be no surprise to those familiar with the arguments, pro and con, concerning objective versus essay tests of achievement. Indeed, similar results for the reliability of direct measures of writing skill have been reported at least since Lindquist (1927) attempted to detect the effects of a laboratory approach to the teaching of freshman composition by using general ratings of the quality of essays. When one compares the statistics in Table 1 with the reliability estimates--typically ranging from .85 to .95 for professionally developed instruments--reported for objective tests of language skills, some of which are clearly less related to writing skill than others, a discouraging picture emerges. On the one hand, the writing assignment, as a direct sample of the type of behavior of interest, has prima facie task validity and is therefore the approach of choice. On the other hand, because of its very nature, the valid task can be measured with only a limited degree of certainty. This dilemma is predictable considering that the typical writing assessment is based on perhaps only one response, whereas the objective test obtains many responses in a much shorter period of time.

Efforts to improve reliability estimates for essay ratings have taken many tacks, including expanding ranges on the rating scales used for evaluation (cf. Coffman, 1971a), articulating in greater detail aspects of the essay to be evaluated (e.g. Moss, Cole and Khampalikit, 1982; Hunt, 1977), and even using computer technology to collect and score essays (Slotnick, 1972; MacDonald et al., 1982; Hawisher, 1987). Although such efforts have met with mixed results, one conclusion follows: technological developments and new scoring protocols or machines cannot reduce the amount of chance error due to variation in individual performance from one occasion to the next. Technology can have no effect on the consistency of behavior. As discussed below, such variation is no doubt influenced by the psychological complexity of the task being performed. Reductions in this source of variation can only be obtained by lengthening the assessment, or preferably, obtaining multiple responses from each individual (perhaps in the mode of discourse most relevant to the local purpose of the assessment). From this point of view, reducing chance errors in direct evaluations of writing skill of an individual comes at the high price of additional time on the part of both writers and readers.

A final note regarding the results in the table relates to the degree of generalizability of the particular results obtained by Breland, et al. In general, they found narrative tasks to produce less reliable scores than expository or persuasive tasks, and found additional responses from students to produce greater increments in reliability than additional readers. Arguing for writing tasks that reflect instructional programs because they will likely result in more reliable student behavior--expository and persuasive writing were more common in the instructional experiences of these students than was narrative--is consistent with the Breland, et al results. So is arguing that under circumstances of limited resources it is more important to include multiple essays than multiple readers. However, local concerns for evaluating writing are best served by attending to the reasons that reliability was enhanced in a particular way in this study, rather than by expecting that the specific results would be the same.

Sources of Systematic Variation in Writing Samples

The foregoing treatment of chance variation in the writing sample was predicated implicitly on an expectation that something akin to a generic skill in handling written language is canonized as an ultimate goal of higher education, if not as a specific goal of particular curricula. The subject of systematic variation in writing samples compels one to confront this notion of a generic ability to write without illusion of consensus among educators, psychologists, or rhetoricians, let alone the faculty committees contemplating assessment. Just as common sense provides sufficient perspective for us to acknowledge the good and bad days had by writers with pen and paper--keyboard and cathode ray tube, as the case may be--and hence the possible lack of consistency even with reliable readers, so too does it urge us to recognize that a novelist is not an essayist, pamphleteer, or writer of copy for the ad in the Sunday magazine section. Good writing takes many forms and in discussing factors that can contribute to systematic variation in responses to essay prompts from one occasion to the next, matters of form are basic to intelligent dialogue and valid interpretation. The validity of inferences based on writing samples is determined principally by the content and genres embedded in the specified writing tasks, the nature of the scoring procedures used and criteria established, and the uses to which scores are put.

Where the design of an assessment in writing departs from the models typically used to describe test development is in the area of scoring criteria and procedures. To provide a detailed review of scoring protocols used by purveyors of direct writing assessment would shift the focus of this essay from principles to practices. However, the manner in which a particular set of criteria is implemented says much about the validity of the scores ultimately derived, so some attention to scoring procedures is needed. Scoring criteria in writing run the gamut from global ratings of overall quality (known as holistic scores) to analytic evaluations of specific characteristics of samples, such as adherence to standard rules in the mechanics of written language and to rhetorical principles of content, style, and organization (Diederich, 1974). Between the holistic and analytic lie procedures that focus the attention of readers on specific attributes of good writing in a given mode of discourse. The primary-trait scoring method and the focused-holistic method (e.g. Hieronymus and Hoover, 1987) are representative examples. Further discussion of scoring methods is provided by Mullis (1984) and Applebee, Langer, and Mullis (1986).

Approaches to Scoring Essays

Table 2 provides a general characterization of the four most common approaches to scoring essays from a direct writing assessment: holistic, primary-trait, analytic, and focused-holistic.

These four approaches share several features, but differ in fundamental respects that affect the type of use to which scores are best suited. All four yield scores on an arbitrary numerical scale with either four or six points in most applications. Most of the approaches make use of exemplary essays for each score point that are used to clarify scoring criteria for the readers, (although one cannot be sure that because a given approach was used, that these "anchor papers" actually constituted part of the training of readers). Each approach may also incorporate verbal descriptions of the "typical" essay deserving a rating of, say, 2 or 5. Scoring approaches differ with respect to the number of scores that are reported, the amount of training usually required to achieve acceptable levels of reading reliability, and the amount of experience that exists with the method used on a large scale.

Table 2. Essential Features of Four Approaches to Essay Scoring

			Writing Assessment	
	Holistic	**Primary-Trait**	**Analytic**	**Focused-Holistic**
Basic Description	essays scored for overall quality rating scale typically ranges from 1–4 or 1–6, with 0 for unscorable paper anchor papers are *sometimes* used to enhance rater agreement	scoring focused on aspects of essay that are 'primary' given the purpose rating scale usually 1–4 anchor papers are integral part of scoring different criteria for different writing tasks	scoring on many dimensions of writing quality (e.g., mechanics, diction, logic, ideas) separate scores reported for each dimension anchor papers used	a combination of holistic, primary-trait, and analytic reports a single score anchor papers and verbal scoring protocols are integral part of method scores are specific to prompt, but reflect features of all good writing
Advantages	rapid training time for raters yields acceptable level inter-rater reliability long record of successful use reliability evidence from many college-level applications	judgments based on more detailed features of discourse consistent with expert views of the writing process	yields potentially useful diagnostic information on strengths and weaknesses	reliability data available from many instances of use compromise between extreme views of scoring and the writing process
Limitations	not useful for diagnostic purposes scores may not be comparable across samples	longer training time for raters extensive development time for prompts and scoring criteria limited experience with technique in large-scale college-level programs	longer training time for raters can be difficult to obtain acceptable inter-rater reliability separate scores often highly correlated	training complicated because approach tries to 'do it all' diagnostic information is limited
Instances of Use	College Board Adv. Placement–English National Assessment of Educational Progress New Jersey Basic Skills Test-Composition ACT-COMP College-Level Academic Skills Test	National Assessment of Educational Progress	Illinois Inventory of Educational Progress Stanford Achievement Test—Writing Sample	many state-wide writing assessments in public schools Iowa Tests of Basic Skills—Writing Supplement Tests of Achievement and Proficiency

The principal distinctions among scoring approaches have to do with how writing skill is defined operationally. The holistic method views the overall impression made by an essay as being greater than the sum of its parts. Thus, a single, global judgment of quality is the appropriate way to evaluate a piece of written work. In addition, the points on a holistic scale may be defined in a norm-referenced manner, that is, in terms of the range of quality shown in the sample of essays being scored. The primary-trait method considers good writing to be related to the purpose of the writing task and establishes criteria accordingly. Thus, a primary trait system has one set of criteria for what constitutes good persuasion, another set for what constitutes good narrative, and other sets that consider the specific purpose implied by the writing assignment. Analytic scoring separates good writing into good organization, good ideas, good sentence structure, and so on. An analytic approach first identifies the components of the written product to be judged and then establishes levels of performance that correspond to points on the score scale chosen. The focused-holistic method (sometimes called modified holistic) is intended as a compromise approach. In it, readers are usually instructed to consider how an essay exhibits qualities common to all good writing as well as how it responds to the particular demands of a given mode of discourse.

It should be clear from Table 2 and the above discussion that most scoring methods are effective at ranking examinee performance in a general way, but of limited value for diagnostic purposes. Primary trait and focused holistic scoring provide some sense of strengths and weaknesses in that criteria are more specifically defined in terms of the essay topic. We know, for example, that a score of 1 on a persuasive essay means the writer was unsuccessful at marshalling evidence to support a position with which his or her audience may disagree. But the primary-trait or focused-holistic score tells us nothing about the degree of command over diction, language mechanics, and the like. Similarly, a score of 1 in a purely holistic paradigm might mean the ideas and organization are reasonable, but the sentence structure and punctuation are unacceptable. In contrast, an analytic approach provides more diagnostic information, but at the price of more extensive training requirements for readers and more complex implementation and score reporting. The needs of placement may be served well by approaches yielding a single overall score, whereas the needs of individual diagnosis and remediation may require more detailed analysis of samples of student writing.

The Writing Process as a Complex Performance

Concerns for the validity of an assessment can go beyond the definition of scoring criteria. Indeed, scoring criteria themselves are defined on the basis of particular beliefs held about the nature of the writing process, the kinds of tasks that promote its development, and the distinctive features of the final product that indicate a writer has command of the components of this process. On these grounds, production-based measures in general, and writing in particular, have attracted recent attention from cognitive psychologists, who interpret a given task like writing an office memo or participating in a simulation exercise from the point of view of the component processes that are assembled and executed by the examinee when engaged with the task.

In the area of expository writing, Hayes and Flower (1986) approach the process of composition as goal-directed behavior that is hierarchically organized in terms of cognitive structure. Writing, in their view, involves three essential processes of planning, generating sentences, and revising. This is analogous to what the creative artist does when painting from line drawings or photographs, what the architect does when drawing up blueprints for a building from specifications given by a client, or what a composer does at the keyboard when writing a new song. The writing trinity is clearly in harmony with the approach to, say, freshman rhetoric that composition instructors have used for years, so from one angle it is legitimate to ask "What's so novel about this cognitive perspective?" One new feature is the language that is used to clarify our sometimes fuzzy intuitive notions of the writing process, notions that instructors use tacitly in organizing lesson plans and that faculty may unwittingly act on when faced with designing an assessment.

In describing production tasks like writing, cognitive psychologists use terms like declarative and procedural knowledge, mental representations of experience that are in some way linked with a strategy for effective communication as an individual plans, generates, and revises written work. Hayes and Flower (1986) contend that strategic knowledge is central to understanding the cognitive demands of a writing assignment. Strategic knowledge is their referent for a combination of three factors: knowing how to define the writing task for oneself with appropriately demanding yet manageable goals; having a large body of high-level procedural knowledge on which to draw; and finally, being able to monitor and direct one's own writing process.

The process orientation to the writing task exemplified here suggests a unique approach that a cognitive psychologist would take to analyzing writing tasks. Much of the effort in understanding this orientation involves clear descriptions of the ways in which novices and experts approach a complex task. With respect to writing, Hayes and Flower (1986) distinguish expert writers from novices primarily on the basis of their planning and revising abilities. They implying that experts at a given task exercise more variety in their capacity to "discover what one wishes to say and to convey one's message through language, syntax, and content that are appropriate for one's audience and purpose," to borrow from a definition of competence in writing given by Odell (1981).

The process approach can only complicate the matter of validly interpreting scores from assessments of writing skill. To the extent that a production-based measure is inseparable from planning and revising, however, we have in the cognitive approach a potential explanation for the variability in performance discussed previously as characteristic of data from writing samples. In this cognitive view, writing is such a complicated process that individual performance will not likely display the consistency that is characteristic of scores on objective tests of writing skill. This line of research is presently moving toward the development of scoring criteria for writing samples that provide a better indication of the cognitive components of the writing process, a development that should prove useful in instructional diagnosis and remediation.

Exemplary Approaches to Assessment and Instruction

Writing samples have formed a part of several well-known tests of language skills for many years, although the majority of these tests have been aimed at admission and placement decisions. Coffman (1971b) describes the development of the College Board achievement test series, whose tests in English and modern foreign languages have included assessments of both writing and listening skills in formats more typically associated with performance assessment. Responses to essays are typically weighted along with scores from objective portions of these instruments to obtain a composite score for use in placement.

Measures of knowledge or achievement in written language that make exclusive use of essays are more difficult to find. For example, Stiggins (1987) refers in passing to the fact that nearly three-fourths of the States are in the process of conducting or developing direct assessments of writing skill statewide. His "Guide to Published Tests of Writing Proficiency" (Stiggins, 1981), however, contains information only on indirect approaches to measurement of writing skill. No doubt much of the effort being expended on writing assessment is being directed toward the design of new instruments that meet local needs. The National Assessment of Educational Progress report on writing (Applebee, Langer and Mullis, 1986) and the teacher's guide for the writing portion of the Iowa Tests of Basic Skills (Hieronymus and Hoover, 1987) describe in detail stages in the development of direct writing assessments for a general public school population. Specific examples of large-scale assessment of writing at the college level include the New Jersey Basic Skills Testing Program (NJBST), which combines an essay score with an objective test to measure English language skills, Florida's College-Level Academic Skills Test (CLAST), which includes an essay in a battery of so-called rising-junior exams, and various tests of the outcomes of general education such as ACT-COMP and ETS's Academic Profile.

Embedded Writing Assessments: Standardized Instruments

As part of a general instrument for measuring outcomes in postsecondary education, the College Outcomes Measurement Program (COMP) (Forrest and Steele, 1982) of the American College Testing Program includes a direct assessment of writing with analytic scores evaluated in terms of attention to audience, organization, and language. The COMP actually assesses both writing and speaking skills in a manner intended to measure the ability to communicate in contexts related to social institutions, science and technology, and the arts. All writing prompts ask the student to address a letter to either a friend, legislator, or administrator that is rated in terms of the three scales mentioned above.

Information about the reliability of COMP's speaking and writing is provided by Steele (1986), who reported correlations between raters--no indication was made regarding whether statistical adjustments were made--that ranged from .87 to .99, with an average between-rater correlation of .94 for writing samples. Correlations between parallel forms of the COMP Writing Assessment were estimated in two separate studies to be .64 and .71, and norms tables with raw-score to percentile rank conversions were given for two reference samples, one of about 4,000 freshmen from 22 four-year colleges and universities and another of about 3,500 seniors at 27 institutions. What is difficult to tell from the available information regarding this assessment is the extent to which reliability estimates might change if determined from other approaches to scoring, and how such changes might affect the normative information provided by the assessment. Given the multiple sources of variation in scores from direct writing assessments, confidence bands for percentile ranks would be a useful addition to scores reported from COMP essays. What is clear from the results reported by Steele (1986) is that the COMP writing scores are similar to scores from other direct assessments of writing skill in that they show greater consistency in the performance of raters than they do in the performance of individual examinees. The implications of this finding for local uses of COMP or other college outcome measures of writing are among the factors to be weighed and considered in practice.

The NJBST includes one essay each year that is scored by raters using the holistic method, which is well-advised in situations where the primary use of scores is for placement or general prediction. A panel of specialists reviews data from pretesting of ten essay topics submitted by panel members before selecting the topic for the assessment in a given year. As there is only one topic per year, the principal reliability concern is the consistency of readers. The use of a six-point holistic scale for essay ratings resulted in a reading reliability coefficient of .88, indicating that trained readers were able to obtain substantial consistency in their evaluations of overall quality of a sample of essays used for estimating reliability (NJBST, 1983). The principal safeguard against year-to-year differences in the difficulty of essay topics, and hence what guarantees comparability of scores across years, is the care taken in the design and selection of prompts.

Embedded Assessments of Writing: Classroom Instruction

Approaches to writing assessment are also embedded in approaches to writing instruction. An insightful example comes from the literature on writing across the curriculum, an instructional orientation whose proponents argue that effective writing instruction (and assessment for that matter) is achieved by emphasizing the importance of writing in all subject areas and by developing activities that engage students in writing to a variety of audiences in a variety of disciplines (Odell, 1980). In a writing across the curriculum approach, the evaluation of writing skill often takes place within the discipline. Examples of such programs in higher education are described by King (1982) and Thaiss (1982). Note that there is nothing unique about this orientation that requires a particular approach to assessment or scoring, or that obviates the need to establish reliability. However, reliability and validity may be arguably enhanced by focusing assessment on the subject matter of a discipline with which the student is familiar.

Practical Considerations

It goes without saying that institutions must necessarily come to terms with a purpose for assessment, select an approach, execute an assessment program, and report results. How does an institution or academic unit approach these concerns in the case of production-based measures like writing? In the case of direct assessment of writing skill, there is no such thing as an off-the-shelf instrument which will yield scores for individuals with a well-estimated degree of chance error. Neither is there a uniform implementation of an approach that will be necessarily valid for the purposes for which scores were sought in a local institution. Field testing of established scoring protocols has not been sufficiently extensive to provide potential users of a given approach with a basis for predicting what levels of reliability might be expected with that approach implemented locally. Reliability and validity of production-based measures are defined through the care and planning that go into an individual implementation and not simply by data reported in a technical manual.

Several themes in this essay provide a basis for discussion of general guidelines in the conduct of a direct writing assessment. The first is grounded in an understanding of the many components of variability in individual scores. To obtain the levels of reliability considered necessary under most circumstances for purposeful individual measurement, assessments of writing skill must strive for multiple ratings of as many instances of performance as are feasible. One essay, read by one or two judges is too brief a behavior sample to be trustworthy. While some may argue that shorter, less reliable measures are adequate for group assessments used, for example, in program evaluation, it remains true that if any meaningful results are to be reported to individuals some degree of reliability needs to be established. An approach in which students actually develop a portfolio of their written work (much like the photographer, graphic artist or computer programmer), begins to approximate the kind of repeated measurements that can assure reliable and valid decisions from a psychometric point of view. Portfolios, however, do not serve the needs of institution-wide or state-mandated assessment.

A second general suggestion derives from the literature regarding performance in different modes of discourse as well as the cognitive orientation to writing tasks that emphasizes the role of procedural and declarative knowledge in writing. To control for mode of discourse and background knowledge, the appraisals of writing skill that are most interpretable and responsive to institutional goals would be those made within the academic major. A general writing assignment may well represent a different cognitive task for the business major and the history major. The response of either as a general indicator of skill in writing may well ignore important explanations of individual differences between their responses. Better would be an approach to assessment that asked the business major to draw on relevant knowledge and experience to analyze and evaluate a hypothetical marketing decision in the format of an office memo, that asked the history major synthesize data from local archives into a brief report on labor union practices of local industry, or that asked a teacher trainee to write a letter home to parents that explained what a child's science curriculum would include during the next two months of the school year. Such tasks would be relevant from the standpoint of instructional or curricular validity and would minimize potential sources of individual differences in performance caused by varying degrees of prior knowledge or familiarity with the particular writing task. In each case, examinees would be prompted with tasks that are appropriately demanding for a given academic major.

Implications for Other Performance-Based Measures

The considerations for the development of reliable and valid direct assessments of writing skill ought to be understood from the broader perspective of performance measurement if the point of view expressed in this essay is to have its intended scope and influence. Berk (1986) defines performance assessment broadly in arguing that it "is the process of gathering data by systematic observation for making decisions about an individual" (p. ix). In this essay, as well as in others in this volume that treat the problems of so-called production-based measures, the writing sample is taken to be a prototypical example of any measure that represents a sample or product of a complex behavior to be evaluated in a systematic fashion. It is not unlike the simulations used in professional schools to determine the effectiveness of management training (Sokol and Oresick, 1986), the structured foreign language interviews used by the foreign service to screen applicants for overseas assignments (Bachman and Palmer, 1981), or even the work sample tests of job performance used for evaluation of training in the military and private industry for decades (Seigel, 1986; Fitzpatrick and Morrison, 1971).

The problems faced in direct writing assessment may well understate the complexities of performance measurement in other professional or vocational contexts. Nevertheless, the critical issues that any institution must address in devising tasks for the measurement of performance, from defining the domain and skills of interest, to designing the materials, to selecting criteria and establishing the reliability of the assessment are ones that require considerable time commitments and dedication from all who are party to the assessment: students, faculty, administration, as well as professional and clerical staff. To be sure, advances in theory of the cognitive processes involved in various types of writing performance and in technology used to conduct assessments will have an impact on how performance data are collected and interpreted. In the meantime, however, it is important to recognize that direct performance assessment is very much of an art from the point of view of psychometric quality. As in all artistic endeavors, one proceeds with the hope of creating beauty knowing full well that it seldom comes on the first try.

References

Applebee, A.N., Langer, J.A. and Mullis, I.V.S. Writing: Trends Across the Decade, 1974-1984. (National Assessment of Educational Progress Rep. No. 15-W-01). Princeton, NJ: Educational Testing Service, 1986.

Bachman, L.F. and Palmer, A.S. "The Construct Validity of the FSI Oral Interview." Language Learning, vol. 31 (1981), pp. 67-86.

Berk, R.A. (ed.). Performance Assessment: Methods and Applications. Baltimore: The Johns Hopkins University Press, 1986.

Breland, H.M. The Direct Assessment of Writing Skill: a Measurement Review. (College Board Report No. 83-6). New York: College Entrance Examination Board, 1983.

Breland, H.M., Camp, R., Jones, R.J., Morris, M.M. and Rock, D.A. Assessing Writing Skill. (Research Monograph No. 11). New York: College Entrance Examination Board, 1987.

Carroll, J.B. "Measurement of Abilities Constructs." In E. Belvin Williams (ed.), Construct Validity in Psychological Measurement. Princeton, NJ: Educational Testing Service, 1980, pp. 23-39.

Coffman, W.E. (1971a). "Essay Exams." In R.L. Thorndike (ed.), Educational Measurement. 2nd ed. Washington, D.C.: American Council on Education, 1971, pp. 271-302.

Coffman, W.E. (1971b). "The Achievement Tests." In Angoff, W.F. (ed.), The College Board Admissions Testing Program. New York: College Entrance Examination Board, 1971, pp. 49-77.

Diederich, P.B. Measuring Growth in English. Urbana, IL: National Council of Teachers of English, 1974.

Ebel, R.L. "Estimation of the Reliability of Ratings." Psychometrika, vol. 16 (1951), pp. 407-424.

Fitzpatrick, R. and Morrison, E.J. "Performance and Product Evaluation." In R.L. Thorndike (ed.), Educational Measurement, 2nd ed. Washington, D.C.: American Council on Education, 1971, pp. 237-270.

Forrest, A. and Steele, J.M. Defining and Measuring General Education Knowledge and Skills. Iowa City, IA: The American College Testing Program, 1982.

Hawisher, G.S. "The Effects of Word Processing on the Revision Strategies of College Freshmen." Research in the Teaching of English, vol. 21 (1987), pp. 145-159.

Hayes, J.R. and Flower, L.S. "Writing Research and the Writer." American Psychologist, vol. 41 (1986), pp. 1106-1113.

Hieronymus, A.N. and Hoover, H.D. Iowa Test of Basic Skills: Writing Supplement Teacher's Guide. Chicago: Riverside Publishing Company, 1987.

Hunt, K. "Early Blooming and Late Blooming Syntactic Structures." In C.R. Cooper and L. Odell (eds.), Evaluating Writing: Describing, Measuring, Judging. Urbana, IL: National Council of Teachers of English, 1977, pp. 91-104.

King, B. "Using Writing in the Mathematics Class: Theory and Practice." In C. Williams Griffin (ed.), Teaching Writing in All Disciplines. San Francisco: Jossey-Bass, 1982, pp. 39-44.

Lindquist, E.F. The Laboratory Method in Freshman English. Unpublished doctoral dissertation, The University of Iowa, 1927.

Lindquist, E.F. "Preliminary Considerations in Objective Test Construction." In E.F. Lindquist (ed.), Educational Measurement. Washington, D.C.: American Council on Education, 1951, pp. 119-158.

Lloyd-Jones, R.J. "Primary Trait Scoring." In C.R. Cooper and L. Odell (eds.), Evaluating Writing: Describing, Measuring, Judging. Urbana, IL: National Council of Teachers of English, 1977, pp. 33-66.

MacDonald, N.H., Frase, L.T., Gingrich, P.S. and Keenan, S.A. "The Writer's Workbench: Computer Aids for Text Analysis." Educational Psychologist, vol. 17, no. 3 (1982), pp. 172-179.

Moss, P.A., Cole, N.S. and Khampalikit, C. "A Comparison of Procedures to Assess Written Language Skills at Grades 4, 7, and 10." Journal of Educational Measurement, vol. 19 (1982), pp. 37-47.

Mullis, I.V.S. "Scoring Direct Writing Assessments: What are the Alternatives?" Educational Measurement: Issues and Practice, vol. 3 (1984), pp. 16-18.

Odell, L. "Defining and Assessing Competence in Writing." In C. R. Cooper (ed.), The Nature and Measurement of Competence in English. Urbana, IL: National Council of Teachers of English, 1981, pp. 95-138.

Odell, L. "The Process of Writing and the Process of Learning." College Composition and Communication, vol. 31 (1980), pp. 42-50.

Ryans, D.G. and Frederiksen, N. "Performance Tests of Educational Achievement." In E.F. Lindquist (ed.), Educational Measurement. Washington, D.C.: American Council on Education, 1951, pp. 455-494.

Quellmalz, E.S. "Writing Skills Assessment." In Berk, pp. 492-508.

Seigel, A.I. "Performance Tests." In Berk, pp. 121-142.

Sokol, M. and Oresick, R. "Managerial Performance Appraisal." In Berk, pp. 376-392.

Steele, J.M. "Assessing Reasoning and Communicating Skills in College." Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1986.

Stiggins, R.J. "Design and Development of Performance Assessments." Educational Measurement: Issues and Practices, vol. 6 (1987), pp. 33-42.

Thaiss, C. "The Virginia Consortium of Writing Programs: A Variety of Practices." In C. Williams Griffin (ed.), Teaching Writing in All Disciplines. San Francisco: Jossey-Bass, 1982, pp. 45-52.

Using the Assessment Center Method to Measure Life Competencies

by William C. Byham

There are a variety of outcomes of higher education that transcend individual course content. These include competency in areas such as decision making, communicating, leading others, and effective social interaction. While these competencies may be covered in individual courses, most are acquired from the variety of academic and nonacademic stimuli to which a student is exposed as part of the entire educational experience.

These outcomes may be considered "life competencies" that are important to the functioning of a mature adult and are required for successful performance in many vocations. For example, the American Assembly of Collegiate Schools of Business (AACSB) determined that the competencies required in business include (AACSB, 1987):

Leadership
Oral Communication/Presentation Skills
Written Communication
Planning and Organizing
Information Gathering and Problem Analysis
Decision Making
Delegation and Control
Self-objectivity
Disposition to Lead

And the Pennsylvania Pre-Teacher Assessment Consortium listed the following competencies as requirements for elementary and secondary school teaching (Millward and Ashton, 1987):

Leadership
Monitoring
Strategic Decision Making
Written Communication
Problem Analysis
Oral Presentation
Initiative
Planning and Organizing
Sensitivity
Oral Communication
Tolerance for Stress
Tactical Decision Making
Innovativeness

Because many life competencies are behavioral, they are difficult to measure with traditional paper-and-pencil instruments which measure knowledge rather than performance skills. These life competencies can best be observed during activities or performance.

The assessment center method has been adopted from industry and government to assess these life competencies in college students. This essay will describe the nature, limitations, and applications of the method in higher education. The term "life competencies" was chosen to describe the types of outcomes usually evaluated by the assessment center method. The methodology also can be used to evaluate specific course or program outcomes as long as they are behavioral (or performance bound) in nature. In fact, this chapter will present examples of the method used in the context of all four major purposes of assessment defined by Millman in this volume.

It should be noted that we will be describing a highly adaptable methodology, not a place or program. An assessment center is a comprehensive, standardized process in which techniques such as situational exercises and job simulations (e.g., discussion groups and presentations) are used to evaluate individuals for various purposes. A number of trained evaluators observe the assessee in the exercises, compare notes, and reach a consensus evaluation.

The key components of the assessment center method are:

1. Organizing the assessment process around target competencies (e.g., Leadership, Analysis).

2. Using a system that assures coverage of all target competencies and uses appropriate inputs from multiple sources.

3. Using past behavior to predict future behavior.

4. Using simulations and other techniques to stimulate behavior to be observed.

5. Having two or more assessors observe and evaluate assessees independently.

6. Having the assessors systematically share and debate their insights into the behavior of those assessed, and relate these insights to each target competency before reaching an overall decision on individual performance, or alternatively, using a computer to combine assessor evaluations using weightings obtained from a panel of experts.

The reliability and validity of the assessment center method for predicting supervisory and managerial success has been well documented (Cohen, Moses and Byham, 1974; Moses and Byham, 1977; Thornton and Byham, 1982; Byham, 1987). There are five uncontaminated, predictive validity studies showing significant results (Bray and Campbell, 1968; Moses, 1973; Moses and Wall, 1975; Hinrich, 1978; and Slivinski, 1979). In these studies, individuals were assessed but no immediate use was made of the data. Later the organizational progress or performance of the assessees was compared to their initial assessments (correlations ranged from .298 to .60). In one study, AT&T was able to predict advancement in the company 25 years after the individuals' initial assessments as new college hires (Howard and Bray, 1988). These five studies are considered to be the best proof of assessment center validity, as the field nature of other studies introduces potential bias into the evaluations.

In the uncontaminated studies, great pains were taken to keep all knowledge of the assessment center findings from the organization, lest some form of bias impact on the research results. In most operational assessment centers, data are used by the organization to aid in initial promotion or placement decisions. Thus, using first promotion as a criterion is inappropriate, though later promotions, turnover, and job performance are more appropriate. A number of studies using these criteria have been conducted by a wide variety of organizations including AT&T, IBM, General Electric, Standard Oil (Ohio), Sears, American Airlines, and the Canadian Public Service Commission.

Another major category of studies evaluated individuals hired or promoted using the assessment center method versus those hired or promoted using another methodology, usually interviews. Seven of nine reported studies show significant differences in job performance. Notwithstanding possible concern over comparability of groups, these studies strongly support the idea that assessment centers are better able to select candidates who are more likely to succeed.

Another way of evaluating the validity of the methodology is by assessing a group of currently good performers and a group of poor performers to see if the assessment center can discriminate accurately between the two. This procedure has been utilized with positive results by a number of organizations including Metropolitan Transit Authority of New York, Detroit Edison, Tennessee Valley Authority, Federal Aviation Administration, American Airlines, Ezaki Gliho Company Ltd. (Japanese ice cream maker), and the State of Massachusetts Vocational Rehabilitation Department. While the majority of this class of validity studies showed positive results, some did not, perhaps reflecting the very difficult criterion problems represented in this type of research.

How an Assessment Center Works

Assessment centers employ a number of techniques to ensure complete coverage of the life competencies sought. Interactive games, leaderless group discussions, role playing exercises and other simulations are used most frequently. These techniques allow participants to engage in life-like situations and display relevant behaviors: decision making, discussions in small groups, and one-to-one interactions.

Individuals are usually assessed in a group. This provides opportunities to observe peer interactions and aids the efficiency of observation. Anywhere from one to twelve people might be observed in a center. For example, a common arrangement consists of six assessees, three assessors, and one program administrator. The low ratio of assessees to assessors (typically 2:1) is important to the assessment center process because it allows close contact and observation and makes multiple evaluations possible. Within an organization, the assessors are generally chosen from among those who do not have direct daily interactions with the assessees.

Observing complex social behavior, integrating the information and making predictions are difficult tasks; therefore, most assessment center programs include extensive assessor training. Their ability to make accurate predictions has been borne out in the literature reviewed by Thornton and Byham (1982) and is supported by findings from other multiple assessment procedures summarized by Cronbach (1970) and Taft (1955, 1959).

Three-Stage Process

The assessment center process can be depicted in three stages. Stage 1 covers the observations and ratings in exercises. Stage 2 includes the reporting of exercise information and the derivation of competency ratings in the staff discussion. Stage 3 encompasses the integration of competency ratings to form a final overall assessment rating.

Figure 1 (Thornton and Byham, 1982) presents a model of the assessment process for one participant observed in a program in which three assessors used four exercises to assess five performance competencies. Stage 1 takes place during and immediately after each exercise when observations and competency ratings are made independently by an assessor.

In the hypothetical example, Assessor A observed the analysis exercise and rated the person 5 on decision making, 4 on oral communication, and 4 on written communication, using a rating scale from 1 (low) to 5 (high). Assessor A also observed and rated performance in the interview simulation. In this exercise, two additional competencies were rated--leadership and use of delegation. Assessor B was the primary observer of this person in the leaderless group discussion, although the other two assessors present were watching other participants in the discussion. Assessor C rated the individual's in-basket performance on all five competencies.

The figure shows that written communication and use of delegation were observed twice, but all other competencies were assessed three or more times. It is desirable to have several

Figure 1.
Assessment Model for an Individual Assessee
(3 assessors, 5 dimensions, 4 exercises)

	Stage 1 Dimension Rating by Exercise			Stage 2 Dimension Ratings Considering the Data					Stage 3 Overall Assessment Ratings			
Exercises (Primary Observer) Dimensions	Asessor A	B	C		Preliminary Assessors A	B	C	Final	Preliminary Assessors A	B	C	Final
Analysis Exercise												
(Assessor A)												
Decision Making	5											
Oral Communication	4											
Written Communication	4			DM:	5	3	1	2				
Leadership Group Discussion												
(Assessor B)												
Decision Making		3		OC:	3	3	3	3				
Oral Communication		3										
Leadership		4										
Interview Simulation				L:	2	4	4	2				
(Assessor A)									2	3	2	2
Decision Making	3											
Oral Communication	3			Del:	2	2	2	2				
Leadership	2											
Use of Delegation	2											
In-basket												
(Assessor C)				WC:	5	2	5	5				
Decision Making			1									
Oral Communication			2									
Leadership			5									
Use of Delegation			2									
Written Communication			5									

"readings" on each competency. Assessors are not asked to evaluate competencies that do not apply to an exercise (e.g., written communication cannot be observed in a group discussion). Stages 2 and 3 take place in the staff discussion when all information is integrated by the assessors. Competency ratings are derived in stage 2. Assessors report observations and ratings from the exercises. Then each assessor independently records his or her preliminary competency ratings. Usually these ratings are displayed on newsprint, a flip chart, or chalkboard for easy examination. In the example, there was clear agreement on oral communication (all 3s) and use of delegation (all 2s). Assessor B rated the person lower than did other raters on written communication, and A rated low on leadership. The assessors disagreed widely on decision making, possibly because the competency was not clearly defined.

During the ensuing discussion, the assessment team, guided by a program administrator, arrives at the final competency ratings. These ratings are consensus judgments, not averages. Many times the team can easily arrive at the final competency ratings (e.g., 3 for oral communication and 2 for use of delegation). At other times, lengthy discussion is necessary. Considering written communication, the group would try to understand why Assessor B gave a 2. In this instance, discussion led to the higher rating. In contrast, subsequent discussion led the other assessors to concur with Assessor A's lower rating for leadership even though initially both scored the assessee above average. The final competency rating for decision making was different from any one of the preliminary assessor ratings.

In stage 3, the staff arrives at the overall assessment rating, defined, for example, in terms of the probability of success as a middle manager or master teacher, or by the more traditional language of overall level of accomplishment. At this stage, preliminary ratings are made independently, posted for examination, and finally consolidated in a consensus discussion. An overall assessment rating is appropriate for selection and evaluation programs, but when assessment is done for diagnostic purposes, stage 3 may be omitted.

Processes of an Assessment Center

Determining Competencies

Assessment centers must be built around competencies, or as many organizations refer to them, dimensions. A competency (dimension) is a description under which behavior to be evaluated can be reliably classified. There are two common sources of competencies: one is from inputs (a sampling of what is being taught to the individual), and the other from outcomes (what the individual needs to be able to do or know after graduation). Current applications of assessment methodology in higher education have primarily used outcomes. The AACSB did not start its "outcomes measurement project" by studying what is taught in business schools. Rather it looked at what makes people successful in business. Similarly, the Pennsylvania Pre-Teacher Assessment Consortium project did not ask what was taught in schools of education: it determined what makes a teacher successful.

In the AACSB project, an extensive review of the professional literature on management success was conducted and then reviewed by committees of business people and researchers specializing in job analysis techniques. Similarly, when the Pennsylvania Pre-Teacher Assessment Consortium wanted to develop an assessment center, it researched the competency literature on teaching first, and then sought input from committees of teachers and administrators (Millward and Ashton, 1987).

When competencies are selected based on outcomes there will usually be a few competencies reflecting unique characteristics of the target group and many competencies that reflect common behaviors required for success in life. These shared competencies include oral communication, interpersonal relations and decision making. The shared competencies will differ among groups in terms of how much is required and the type and amount of inputs and outputs. Decision making for a mechanical engineer is quite different from that of a physician. Communication for a teacher is different from that of an accountant. These differences must be reflected in the detailed definition of each competency that is used in assessor training and around which the assessment center data integration is based.

Competencies (dimensions) defined from inputs are developed from a comprehensive survey of the curriculum, educational experiences, and extracurricular experiences. This task is usually performed by multiple committees which survey the offerings, develop tentative competencies, and then rate and rank competencies until a final list is developed.

One should note that the most common way of determining competencies in government and industry is through a formal job analysis of target positions. Many procedures have been worked out to accomplish these analyses (Drauden and Peterson, 1974; Fine, 1986; McCormick, 1979; Flanagan, 1954). However, because educational institutions are interested in broad groupings of outcomes--not a specific job--the traditional, formal methods seldom suffice.

No matter what the source of the original lists of competencies, as much participation and as much agreement as possible must be obtained through direct and indirect (questionnaires) consultation. But consensus is not the only criterion of a good competency--good definitions are extremely important. For an assessment center to work, trained assessors must be able to agree that observed behavior fits under a specific competency. This consensus is much harder to attain than one might think, and requires great precision in developing definitions. Competencies that are defined "casually" cause confusion, decrease the reliability of the methodology, and waste time in the assessment center process.

A secondary issue in establishing competencies is determining the scale to be used in evaluating the competencies. A rating scale such as the following is often used:

5	A great deal of the competency was shown (excellent).
4	Quite a lot was shown.
3	A moderate amount was shown (average).
2	Only a small amount was shown.
1	Very little was shown or the competency was not shown at all (poor).
0	No opportunity existed to observe the competency .

Of course the terms must be defined. Usually the definition is handled by committees of experts who take outcomes of the assessment center exercises and indicate what behavior would be associated with each rating. A statistical technique called "policy capturing" can be used. In this technique, experts are given behavioral examples (reflecting a wide range of performance) which they sort into the designated categories. Their decision-making strategy requires them to note how they weight the variables that determine the sorting. Using these data, a formula is developed to help less experienced assessors make the same decisions.

Choosing Assessment Simulations

Simulations used in academic assessment centers come in a wide variety of forms. Some typical simulations are:

- o Oral Presentation--The assessee is asked to make an oral presentation (content is provided, if appropriate). The assessee also may be asked to answer questions afterwards.

- o Assigned Group Leadership--The assessee is assigned to lead a group which must resolve a problem. The other members of the group are role players. The assessee must present the problem, conduct the meeting, and achieve some consensus.

- o Group Membership/Leadership (assigned positions)--In a leaderless group discussion, each assessee in the group is given a point of view to present and champion. Assessees are evaluated on their ability to present their point of view, persuade others and lead the group.

- Group Membership/Leadership (non-assigned positions)--A leaderless group of assessees is given a problem to solve. The assessees must organize their efforts, discuss options and make a decision. Leadership must emerge from the group.

- Individual Leadership--The assessee is put in a position where he or she must influence another person to change his or her behavior. After studying background data on the situation and the person, the assessee conducts a one-on-one meeting with a role player.

- Fact Finding and Decision Making--The assessee is provided information on a decision that must be made. More information can be obtained through questioning or research (in a library, or using data bases). After the assessee has obtained all the information he or she deems appropriate, the assessee must make and defend the decision.

- Analysis and Decision Making--The assessee is provided with extensive analytical data which must be organized and analyzed. A decision must be made and communicated in oral or written form.

- Planning--The assessee must develop a plan to accomplish a major activity such as the installation of a new computer system. The assessee must present and defend the plan.

- In-basket--The assessee plays the role of a newly appointed supervisor who must handle the memos, letters and reports contained in an in-basket. Decision making, initiative, decisiveness, delegation and control are some of the areas usually assessed.

Absent from these illustrations are performance measures used in evaluations of writing, art and music. These are centered on particular academic disciplines and are covered by Dunbar and Adelman elsewhere in this volume.

The procedure that determines target competencies also guides the selection of appropriate exercises. The exercises define the competencies as much as the definitions do. For example, if one is trying to assess the competency "planning," then day-to-day planning, long-range planning, or life planning might be considered. The nature of the exercise in which the competency is assessed focuses the assessment on the type of planning that has been defined in the job analysis. Actually, multiple relationships are involved.

It is not easy for a potential user to differentiate a good simulation from a poor one. The success of a simulation depends on its ability to elicit the required behaviors that are to be observed. It's easy to develop a situation where a candidate must interact with someone else or to create an in-basket exercise that contains items an individual might find in a real in-basket. These might look like assessment center simulations, but they may not be effective. In subtle ways, they might "telegraph" the right answer; they might allow the individual to get "off the hook" by avoiding committing himself or herself; and most often, they may rely too much on chance to elicit the desired behavior (Crooks, 1977).

A good assessment center simulation does not "hope" that someone will show leadership, delegate assignments, or be sensitive to others. A good simulation puts assessees in a situation where they must do something or be rated poorly in the simulation. It uses "distractors" appropriately and doesn't use tricks. To mislead the assessees intentionally would be to influence the reliability of the evaluation. Spending time and effort on developing good simulations pays off immensely in the efficiency of the assessment center.

The specific content of a simulation (e.g., topic of a group discussion exercise, issues to be explored in an in-basket exercise) should be chosen to provide an equal challenge to all participants. The observation of behavior will be distorted (and hence rendered unreliable) if different content knowledge produces different results. Making exercises "content-fair" is much easier than it might seem. Group discussions or fact finding tasks should focus on topics not covered in any course. Good topics include any future-oriented issues or people issues. In-baskets and other exercises are often set in hypothetical countries to remove specific content references.

The difficulty level of an exercise must be carefully equated to the performance criterion that is to be predicted. For some competencies, a linear relationship between achievement and success can be assumed (e.g., decision making). For other competencies, a minimum level of skill is all that is necessary. An excess of the skill will not make one much more successful in most areas of life. Written communication is such a competency for many college graduates.

If a linear relationship with success is assumed, then an exercise must distribute participants on the competency. The difficulty level should be such that participants are not bunched at the high or low end of the distribution. If a minimum acceptable level of competency is required, then this benchmark can be set by asking experts to define behavior just above and below the minimum. The exercise is then developed to highlight that decision point. It is very important that assessor forms and training focus attention on this critical benchmark.

If unique assessment exercises are developed, they must be tested thoroughly to assure that the required competencies are elicited. Instructions for participants and observation forms for assessors must be checked to ensure clarity and reliability. Model reports and other materials should be provided to help assessors develop reliable judgments on the rating scales for each competency.

Assessor observation forms differ according to the amount of assessor training provided and how frequently assessors are used. Forms must be more detailed and precise when training time is restricted or when assessors are used only occasionally. Inexperienced assessors usually prefer detailed forms, whereas highly trained and experienced assessors find that any form other than the most general slows them down.

The objective of the observation form is to increase the reliability of judgement. Reliabilities of .8 and higher are the goal of most assessment centers (Thornton and Byham, 1982). The requisite level of reliability depends on the purpose of the assessment center, with greater reliability is required for individual assessment than for institutional evaluation. However, in all cases, reliability needs to be high to produce useful validities.

Selecting Assessors

Assessor teams usually consist of three to six assessors (teams of three are the most common). An individual assessee is observed by one assessor in each exercise. The assessors do not communicate with each other about a participant's performance until the integration meeting. There are some exceptions under which teams of assessors observe every individual in every exercise. Obviously, this drastically increases the cost of the assessment center. On the other hand, reliability is enhanced by having multiple observers. Is the increase in staff or staff work load worth it? Most organizations don't think so. They feel that sufficient reliability can be obtained from a combination of well-crafted assessor observation forms and effective assessor training.

In almost all assessment center applications in business and government, assessors are drawn from management in the organization. Usually assessors hold positions at least two levels above the individuals being assessed. Several universities have attempted to use faculty as assessors--particularly in business schools--but with little long-term success. Alverno College uses a combination of faculty and volunteers from the community, and reports that the volunteers not only relieve the faculty of the time-consuming responsibility, but also provide a more objective judgment. In addition, the use of external assessors seems to provide good public relations, as it involves more people with the college. It also helps the volunteers observe the quality of Alverno graduates.

Professional assessors from firms specializing in selection or assessment can be used to evaluate samples of assessees' behavior which have been collected using paper-and-pencil instruments or videotape. For example, the United States Foreign Service Examinations include an in-basket exercise that is evaluated by a team of specialists at the Educational Testing Service. The written and videotaped exercises used in assessment center projects for the AACSB, University of Pittsburgh, University of California at Berkeley, Florida Atlantic University, and several other programs, were evaluated by a team of specialists at Development Dimensions International.

In the AACSB outcomes measurement project referred to above, there was no choice in the matter of evaluation logistics. Because of the relatively small number of students evaluated in each institution, it would have been extremely difficult to recruit and train assessors at each institution; and because the project also sought interinstitutional comparisons, it would have been difficult to develop common standards among assessors drawn from the different institutions. After all, one criterion in selecting assessors is their freedom (both real and perceived) from bias. In the AACSB situation, local assessors might not have been able to exclude personal knowledge of individuals from their assessment observations, and the public might question whether institutions assessing themselves can provide fair and accurate judgments.

Training Assessors

Adequate and complete assessor training is the key to reliable assessment judgments. Basically, an assessor must know how to recognize, record, categorize, and evaluate behavior. Thus, minimal assessor training involves:

- Thoroughly acquainting assessors with the competencies to be evaluated;
- Allowing assessors to observe or participate in the assessment exercises;
- Illustrating appropriate behavior for each point on the rating scale; and
- Allowing assessors to practice observing individuals participating in the exercise and to share their ratings and reasons for the ratings.

Most assessor training programs involve practice observing the behavior of sample participants presented on videotape. Assessors watch videos of individuals participating in exercises and get to practice their skills. The use of videos allows consistency of training stimuli from one assessor training program to another and develops consistency of judgment as assessors compare their competency ratings with each other and with other groups.

The proficiency of assessors should be determined before they are used in an actual assessment center. Usually this is done through a practical examination at the end of the assessor training program. The assessors evaluate an individual on videotape and those evaluations are compared against a standard.

Administering Assessment Centers

Usually assessment centers are more difficult to administer than typical paper-and-pencil instruments. However, some assessment exercises can be administered in a manner similar to the administration of standardized tests. For example, a large number of people can be brought together and given an in-basket exercise, or a planning or analysis exercise. Experimentation is currently underway with methodology to administer a form of interpersonal exercise to large groups through the use of videotape stimuli and paper-and-pencil responses.

The majority of assessment center exercises require one-to-one or small group activities which, in turn, require extensive scheduling of assessees, role players and assessors. If a role player is needed, he or she must be scheduled without conflicts, and must have proper training. For group exercises, the group must be assembled; and getting students to show up for volunteer activities is not easy.

Other unique problems come about when group exercises are involved. If an exercise is designed for six people it should be used for six people, not four. The drop-out rate of volunteers causes some administrators to schedule eight students to ensure that six will show up.

The administrative requirements of running an assessment center consisting of three or four exercises should not be underestimated. However, the AACSB project proved that individuals with no special training can do it. In the AACSB study, individuals on various campuses were sent complete directions on how to set up and run the assessment center, and did so with only limited phone consultation.

Providing Feedback to Participants

When the purpose of an assessment is program evaluation, the major method of recruiting student volunteers to participate in an assessment center has been to offer a personal benefit in the form of performance feedback. Providing feedback causes its own administrative problems. A personalized written report and appropriate guidance for interpretation must be provided. In addition, an institution usually will want to provide some kind of personal counseling along with the written feedback. Usually students are given written feedback on their performance and told that counselors are available to discuss the results or development implications. The Katz Graduate School of Business at the University of Pittsburgh has been assessing students for four years in connection with various research projects. They have found that five percent of students request additional information on their performance in the assessment center.

Sampling

As discussed elsewhere in this volume, appropriate sampling is very important if the data are to be generalized to a larger student population. When an institution is drawing a true sample, it cannot use volunteers. An expensive way around this problem is to ask for volunteers and then draw from the assessed volunteers a sample representative of the entire student population. This method still results in bias (i.e., only the best students might volunteer), and causes the institution to assess more people than are needed.

For most institutions, a more appealing alternative is to define a stratified, random sample of individuals in an appropriate manner and make sure they're evaluated. This means the evaluation must be made as appealing as possible so that students will participate of their own volition. Further, it implies intensive follow-up of "no show" students and offering convenient reassessment opportunities. Fortunately, it is not difficult to recruit students. It's much more interesting to take part in a group discussion than to answer 100 multiple choice questions.

Security

When the purpose of the assessment is certification or placement, providing appropriate assessment instrument security is always an important issue. Paper-and-pencil instruments such as the in-basket exercise or an analysis exercise can be handled as would a test or questionnaire. They can be administered to large groups over a relatively short period of time. Close control of the assessment material can be maintained.

Security for interactive and presentation exercises is more difficult. Because of the administration challenges noted previously, students often must be assessed over several days or weeks. Thus it is possible for them to share the content of the exercises with others. This problem can be avoided by having parallel forms for the exercises so a student would never know which situation he or she might face, such as in an interactive simulation. Even so, knowing the content of exercises will not help most people in most assessment situations. For example, knowing the topic of a leaderless group discussion will not make a person a leader.

Pros and Cons of the Assessment Center Method

Assessment center methodology has many advantages when compared to other methods of outcome evaluation. These are well documented in professional literature (Thornton and Byham, 1982). Advantages include:

1. Less faking--Participants in assessment center exercises actually perform, not merely explain how they would perform.

2. Content validity--The exercises can be germane to academic content taught, "real world" content, or content important to areas in which the individual might be employed.

3. Face validity--The exercises are "real" to the participants. In the AACSB study, participants said that the exercises were realistic and valuable. Face validity is very important when subjects must volunteer a day or more of their time.

4. Reliability--High reliability across assessor groups makes the methodology fair to all participants no matter when or where they go through the assessment center.

5. Job preview--The exercises provide an opportunity to experience "real world" activities. In the AACSB study, students reported unfamiliarity with the assessment center exercises, but saw them as relevant to the kinds of things they would be doing after graduation.

6. Limited racial or sexual bias--Extensive studies have shown the assessment center method has less adverse impact on minority or gender groups than almost any other evaluation technique (Byham, 1986).

Disadvantages include:

1. Cost--The major cost of the assessment center method is assessor and administrator time. In general, colleges and universities have had little success getting faculty to volunteer their services as assessors on an ongoing basis. Most of the larger, ongoing applications use assessors from outside the organization or provide special pay for internal assessors.

2. Administrative complexity--Assessment center exercises are more difficult to administer than traditional paper-and-pencil tests.

3. Assessor training--Training must be provided for assessors regardless of their academic rank or field of degree. Special skills are involved.

4. Lack of national norms--The AACSB research offers the only case in which national norms are currently available. They are available for the AACSB study because a single group of highly trained people drawn from outside the institutions conducted all the evaluations.

Overcoming Assessment Center Disadvantages: Videotaping

The videotape machine and the computer seem to offer the best solutions for overcoming the disadvantages of assessment centers.

Many assessment center exercises require one or more role players (e.g., an interaction simulation requires a role player to be a subordinate or a peer whom the assessee has to influence). These role players must be recruited and trained. Later, they must be scheduled to play their roles at a particular time and place. A well-prepared videotape can replace role players. The assessee responds verbally to the video role player as an assessor takes notes or records the response on audiotape. The best use of videotape technology for this purpose is illustrated in the assessment center conducted by The Pennsylvania Pre-Teacher Assessment Consortium. In one of the exercises, student assessees play the role of a teacher who must respond to observed interactions in the classroom. The students write down what they would say or do. This procedure allows for the administration of the exercise to large groups. Videotaping assessment center performance can have a major impact on the disadvantages noted previously in the following ways:

> Cost--Videotaping performance renders the role of an assessor more attractive. Thus, faculty are more likely to volunteer. Professors and professionals can view the videotapes of assessee performance at their convenience, and that, too, is an inducement to volunteer. By using videotape, an assessor can rate four individuals per hour as opposed to one or two in a live situation. The increased efficiency far outweighs the cost of tape, equipment, and transportation of tapes.
>
> Administrative Complexity--Use of videotape technology makes assessment centers easier to administer because it lessens the problems of scheduling assessors and assessees to be at the same place at the same time. Also, assessment center exercises can be administered in a "batch mode." For example, all the individuals going through a decision making simulation can be processed together. This would not be possible with live

assessors because they need time between exercises to write reports and get ready for the next exercise.

Assessor Training--Assessor training is minimized by using videotape technology. In a typical assessment center, each assessor has to be able to evaluate every exercise and, thus, must be trained in each. When videotape technology is used, assessors can specialize, requiring in-depth training only on their assigned exercises. Of course, assessors need an overview of the other exercises included in the assessment center so they can compare data effectively.

Lack of National Norms--If the same group of individuals evaluates assessment center exercises, then national, regional, or institutional norms can be developed. This otherwise elusive goal can be reached if the videotaped assessments are drawn from a single location.

There are additional advantages of videotaping exercises with respect to reliability and bias. The reliability of the assessment center method is usually established during assessor training. The use of videotape technology makes it easy to obtain reliability during actual assessment. Multiple evaluations can be obtained on the same assessee, and the ongoing reliability can be monitored through overlapping assessment (e.g., every tenth assessee can receive a second evaluation). As for bias, the assessor receives only a videotape with a code number, and knows nothing about the academic or personal background of the assessee, or whether the assessee is part of a pre- or post-test. This is an important advantage in terms of overcoming real or perceived biases.

Applying the Assessment Center Method in Higher Education

In the first essay of this volume, Millman identified four main purposes of assessment: individual placement; individual certification; course and program evaluation; and evaluation of the institution. Courses, programs, academic sub-units, or entire colleges certainly can be evaluated by comparing the average performance of a representative sample of students in appropriate assessment center measures to either the performance of similar samples taken at different times; or similar samples from different institutions; or to a pre-determined criterion. The principal interest of higher education in the method lies in its value when the individual is the unit of analysis. Thus, the following sections describe special considerations and exemplars in using the method when the purposes of assessment are certification, selection, and career counseling.

Certification

When the assessment center method is used to take a significant academic action such as assigning grades or determining graduation, the target competencies must come from the course work provided to the students. It would be unfair to teach one thing and to evaluate students on another--even if the latter is important to job or life success. Ideally, both the course content and the competencies measured would reflect the criteria for success.

Exercises must also reflect the methodology used in the curriculum. It is not valid to evaluate students using group discussion methodology if they haven't been exposed to this or similar interactive techniques in the classroom. If this guideline is not followed, a "method effect" may result that will confound the intended purpose of the evaluation.

Placement/Selection

An infrequent but potentially useful application of the assessment center method in colleges and universities is for selection of students into academic programs. Under the selection rubric, the criteria for performance (competencies) must reflect the content of the program to which the students will be exposed. The methodology of assessment (simulations) also should be compatible with course methodology. Reliability and fairness are important characteristics.

Only one institution of higher education is using the assessment center method for selecting students into programs. That institution is the Moray House College of Education, Edinburgh, Scotland, which introduced the first assessment center procedure for selecting Bachelor of Education candidates in 1984. In 1985 the procedure was revised, and in 1986 was used to select post-graduate candidates.

Four levels of performance for each of six competencies were described, and formed the rating scale used by assessors to evaluate:

1. A written exercise that required participants to reconcile conflicting evidence about a situation.

2. A "practical teaching task" that required a candidate to study some nonsense words and then teach the words to another candidate. The teaching was followed by an assessment of learners' understanding of the words they had been taught.

3. A 30-minute "employment type" interview that involved one candidate and two assessors.

Up to 96 candidates were assessed in a single day by the college staff and by teachers from neighboring educational authorities. One day of assessor training was provided.

The British method of operating an assessment center differs from the American method in the way exercise observations are processed. British assessors conduct a preliminary review of candidates in the middle of the assessment center process. They then adjust the remaining exercises to elicit specific information on those candidates who seem to be on the borderline between acceptance and rejection.

The Moray House College of Education has validated its selection system with mixed results. Their criteria were end-of-course performance on 12 characteristics measured by grades (with methods classes getting the most weight) and student teaching evaluations. When selection competencies were related to end of course performance, only some of the competencies showed validity at the .05 confidence level. Oral communication, commitment, interpersonal skills, and depth of character correlated with placement grades, for example, while practical teaching ability correlated with none of the performance measures (Wilson, 1987).

Career/Development Planning

In order to engage in career planning, a student needs to understand the competencies required for success in the position of interest and to have an accurate insight about his or her performance level in each competency. In this application of the method, target assessment center competencies should thus focus on what is required in a career. Often an assessment center for career development planning will deliberately avoid assessing competencies for which a student could normally obtain feedback through course grades, and will focus exclusively on other areas such as interpersonal skills.

Quality of feedback to assessees is particularly important in the career development application. Assessees must understand and believe in both the accuracy of the competencies and the evaluation they receive in each. In other words, they must accept the criteria on which they are being evaluated and the accuracy of the measurement method.

While reliability is always important, it is not as crucial in career/development assessment centers. The important outcome is what the assessee takes away in the form of actionable insights. The data from the assessment center are usually combined with other data (course and other achievements) during the career discussion, thus providing a form of checks and balances.

Colleges have been relatively slow in utilizing assessment center technology for career planning, and most of these applications lasted only a short time. Cost, administrative complexity and difficulty in recruiting assessors seem to be the major deterrents.

For example, Baylor University created a program to provide self-development and career planning insights to twelve MBA students per year in its Hankamer School of Business (Williams, 1977; Williams and Longenecker, 1976). Twenty to twenty-five competencies relating to managerial success were assessed using an in-basket exercise, a series of group exercises, an analysis and presentation exercise, and paper-and-pencil tests. The competencies and the evaluation instruments were modified from year to year. The assessment center was conducted at a site away from the university, with faculty members serving as assessors. Program participants were given detailed feedback on the results of the assessment center. Then they participated in a seminar and counseling program involving self-analysis, development of personal goals, information on career plans, and behavior modification where appropriate. The program withered away by 1981, though, because of the enormous amount of time required of participating professors (J.C. Williams, personal communication, 1981). A similar program at the School of Business at Stanford University also lasted but five years.

In 1977, the U.S. Air Force, Air University, Squadron Officers School designed a diagnostic assessment center to measure junior officers' strengths and weaknesses for future development purposes. Approximately 100 assessors were trained simultaneously (via closed circuit television) to evaluate 800 students every 11 weeks. The results of the assessment center were given to the students and, if they desired, a special training program was designed to address weaknesses such as planning and organizing skills, leadership, and communication skills. The results were never included in an official record. The program was dropped in 1979 when the classes were reduced to 8 weeks in length and faculty time was severely limited (C. Austin, personal communication, September 18, 1986).

Course, Program and Institutional Evaluation

A few technical caveats should be noted when the unit of analysis is the course, program, or institution. If assessment center results are used for an administrative purpose such as allocating funds, reliability and freedom from bias are essential. Assessors usually cannot be drawn from the institution. If each institution conducted its own assessment, it would be very hard to obtain the necessary reliability of evaluation across institutions. On the other hand, if assessment center results will be used for institutional self-study, reliability and freedom from bias are less important. Often the acceptance that results from the mere fact of self-assessment outweighs any loss in reliability or freedom from bias.

Conclusion

Developing "life competencies" is one of the goals of most institutions of higher learning. They are defined in college catalogs and discussed when college administrators speak to prospective students and their parents. These life competencies are legitimate goals for institutional assessment and the assessment center method is an effective method for evaluating them. The assessment center method is also effective in measuring specific course or program outcomes. The method has been used for individual placement and certification, course, program, and institutional evaluation.

The assessment center method is an additional tool available to an institution to measure its own effectiveness. With more than 25 years of research into the assessment center method, it is not an experimental procedure. It is a proven, valid, and reliable procedure for assessing many areas difficult to evaluate with other measures, and is particularly fair to all gender and racial groups.

The assessment center method has been used successfully by a number of educational institutions mainly to evaluate outcomes from professional schools. These applications have proven that the method can be made cost effective and practical. Because many of the competencies evaluated are similar to those one would expect from an undergraduate program, the applications also indicate that more generic life competencies such as interpersonal skills and practical decision making skills resulting from undergraduate programs can be measured.

Few institutions would use the assessment center method exclusively. Rather, the method would most likely be used in conjunction with paper-and-pencil instruments to measure specific course or program outcomes. Both methodologies are appropriate for different subsets of competencies. An institution should first determine the competencies to be assessed and then determine the appropriate methodology. It should not start with a methodology and look for competencies to evaluate.

Developing and administering an assessment center is not easy. It involves much more than having faculty members observe students going through simulations. Starting and administering an assessment center is a highly technical process that requires considerable planning, material development and training skills. It should not be attempted by anyone who has not studied and experienced the method.

New developments in assessment center methodology--such as using videotape to record participant behaviors--have produced time and cost savings and have made possible larger norm groups. The method now is at a cost level that renders it a practical consideration for most institutions. Further developments such as the use of computer-controlled interactive video should produce further cost savings and administrative simplifications.

References

Alverno College Faculty Assessment at Alverno College. Milwaukee, WI: Alverno Productions, 1985.

American Assembly of Collegiate Schools of Business Outcome Measurement Project, Phase III Report. St. Louis: Author, 1987.

Berk, R.A. (ed.) Performance Assessment: Methods and Applications. Baltimore: The Johns Hopkins University Press, 1986.

Bray, D.W. and Campbell, R.J. "Selection of Salesmen by Means of an Assessment Center." Journal of Applied Psychology, vol. 52 (1968), pp. 36-41.

Byham, W.C. Applying a Systems Approach to Personnel Activities. (Monograph IX). Pittsburgh, PA: Development Dimensions International Press, 1987.

Byham, W.C. and Robinson, J.C. "Interaction Modeling: A New Concept in Supervisory Training." Training and Development, vol. 30 (1976), pp. 20-33.

Byham, W.C. The Assessment Center Method and Methodology: New Applications and Technologies. (Monograph VIII). Pittsburgh, PA: Development Dimensions International Press, 1986.

Cohen, B.M., Moses, J.L., and Byham, W.C. The Validity of Assessment Centers: A Literature Review. Pittsburgh, PA: Development Dimensions International, 1974.

Cronbach, L.J. Essentials of Psychological Testing. New York: Harper and Row, 1970.

Crooks, L.A. "The Selection and Development of Assessment Center Techniques." In J.L. Moses and W.C. Byham (eds.), Applying the Assessment Center Methods. Elmsford, NY: Pergamon Press Inc., 1977, pp. 69-87.

Drauden, G.M. and Peterson, N.G. A Domain Sampling Approach to Job Analysis. St. Paul, MN: Test Validation Center, 1974.

Fine, S.A. "Job Analysis," in R.A. Berk, pp. 53-81.

Flanagan, J.C. "The Critical Incident Technique." Psychological Bulletin, vol. 51 (1954), pp. 327-349.

Hinrichs, J.R. "An Eight-Year Follow-Up of a Management Assessment Center," Journal of Applied Psychology, vol. 63 (1978), pp. 596-601.

Howard, A. and Bray, D. Managerial Lives in Transition: Advancing Age and Changing Times. New York: Guilford Press, 1988.

Loacker, G., Cromwell, L., and O'Brien, K. "Assessment in Higher Education: To Serve the Learner." In Adelman, C. (ed.) Assessment in Higher Education: Issues and Contexts. Washington, D.C.: U.S. Department of Education, 1986, pp. 47-62.

McCormick, E.J. Job Analysis: Methods and Applications. New York: AMACOM Books, 1979.

Mentkowski, M. and Loacker, G. "Assessing and Validating the Outcomes of College." In Ewell, P. (ed.), Assessing Educational Outcomes (New Directions for Institutional Research, no.47). San Francisco: Jossey-Bass, 1985, pp. 47-64.

Mentkowski, M. and Doherty, A. Careering After College: Establishing the Abilities Learned in College for Later Careering and Professional Performance. Milwaukee: Alverno Productions, 1983, 1987.

Millward, R.D. and Ashton, B. "Assessing Teaching Skills Prior to College Graduation," Florida Journal of Teacher Education, vol. 4 (1987), pp. 39-44.

Millward, R.D. Pre-teacher Assessment: Development, Implementation and Follow-up. Paper presented at the International Seminar on Teacher Assessment, Edinburgh, Scotland, 1987.

Moses, J.L. "The Development of an Assessment Center for the Early Identification of Supervisory Potential," Personnel Psychology, vol. 26 (1973), pp. 569-580.

Moses, J.L. and Byham, W.C. Applying the Assessment Center Method. Elmsford, NY: Pergamon Press, 1977.

Moses, J.L. and Wall, S. "Pre-Hire Assessment: a Validity Study of a New Approach for Hiring College Graduates," Assessment and Development, vol. 2, no. 2 (1975), p. 11.

Mudd, J.O. "Assessment for Learning in Legal Education." Paper delivered at the Annual Meeting of the American Educational Research Association, San Francisco, CA, 1986.

Read, J. "A Degree by Any Other Name: the Alverno Program." In F. Hughes and O. Jills, (eds.), Formulating Policy in Post-secondary Education: the Search for Alternatives. Washington, D.C.: American Council on Education, 1975, pp. 214-226.

Schaab, N.A. and Byham, W.C. Effectiveness of Computer-based Training/Interactive Video Versus Live Training. Pittsburgh, PA: Development Dimensions International Press, 1988.

Slivinski, L.W. Identification of Senior Executive Potential: Development and Implementation of an Assessment Center. Ottawa: Public Service Commission, 1979.

Stillman, P., et al. "Six Years of Experience Using Patient Instructors to Teach Interviewing Skills," Journal of Medical Education, vol. 58 (1983), pp. 941-945.

Stillman, P. and Swanson, D. "Ensuring the Clinical Competence of Medical School Graduates Through Standardized Patients," Archives of Internal Medicine, vol. 147 (1987), pp. 1049-1052.

Taft, R. "The Ability to Judge People," Psychological Bulletin, vol. 52 (1955), pp. 1- 23.

Taft, R. "Multiple Methods of Personality Assessment," Psychological Bulletin, vol. 56 (1959), pp. 333-352.

Thornton, G.C., III and Byham, W.C. Assessment Centers and Managerial Performance. New York: Academic Press, 1982.

Williams, J.C. "Systematic Career Assessment/Planning for MBA Students." Paper presented at the 37th annual meeting of the Academy of Management, Orlando, FL, 1977.

Williams, J.C. and Longenecker, J.G. Non-traditional Dimensions of Managers of Education in Academia. Paper presented at the 36th annual meeting of the Academy of Management, Kansas City, MO, 1976.

Williams, R., et al. "Direct Standardized Assessment of Clinical Competence." Medical Education. (in press).

Wilson, J.D. "Selection and Grading in Post Graduate Primary Initial Teacher Training in Scotland." Paper presented at the International Seminar on Teacher Assessment, Edinburgh, Scotland, 1987.

Wilson, J.D. Criteria of Teacher Selection (CATS). Project Research Summary. Research report available from Moray House College of Education, University of Edinburgh, Edinburgh, Scotland, 1985.

Conclusion:
Metaphors and Other Guidances in Higher Education Assessment

by Clifford Adelman

In 1985, when the current movement for assessment in U.S. higher education was in its infancy, I wrote a series of polemics on its behalf for delivery to a variety of skeptical audiences. Each of these polemics began with the same extended metaphor. After editing the work of my colleagues (and hence, paying very close attention to what they were saying), I think I finally understand what the metaphor is really about: the way we ask questions of students, and ultimately, the relationships among the stuff of learning, instruction, and assessment. After the weighty pages that have preceded this essay, I would like to take editor's license, and use that story for the last time---minus the one-liners that inevitably occur in oral delivery:

The History Examination

> When you returned to college in September--of 1571--I am sure you recall reading in the papers that approximately 300 galleys of the fleet of the Turkish Empire had been dawdling away their time around the Adriatic Sea, pillaging and sacking various towns on the Dalmatian coast, various islands north of Corfu, and even various shore installations in the Gulf of Venice. All those lovely watering holes of your vacation lay in ashes and smoke.
>
> The Turkish fleet soon tired of its sport. Drawing on the textbook from Cartography 104, the admirals figured out that the Adriatic could be a trap, moved south to the Ionian Sea, picking up biscuits from their advanced bases, and waited to see how a strange group of allies called "The League" would respond to this general provocation and mischief.
>
> "The League" was a somewhat contentious club that had been formed in the course of strenuous negotiations conducted through the good offices of the Pope. Habitually wary of each other, the cardinals and emissaries of kingdoms and states, major and minor, were united by the Pope's appeal to both dim memories of past Crusades and the palpable strains of present economic interests. They were ready to drive the infidels from their shores and spheres of influence. Thus, the League's own armada, consisting of only 200 ships but speaking a half-dozen languages, set sail from Messina on September 16 to find out where the Turks had gone.
>
> As fortune (and history) would have it, the League's first stop was Corfu, where the local intelligence apparatus pin-pointed the location of the infidels as the Gulf of

Lepanto. Not entirely idle themselves, the Turks simultaneously received information concerning the arrival of their late summer guests. What a lovely time of year for an outing!

And, indeed, the two outings found each other at sunrise on October 7. The meeting, however, was hardly romantic. The Turks may have passed Cartography 104, but didn't make it through the sequence of 105. That is, while they correctly analyzed the Adriatic as a trap, they failed to offer a similar analysis of the Gulf of Lepanto. Admiral Don John strung his fleet across the mouth of the Gulf, and placed the heavily armed though half-mobile Venetian galleases in the front line. The rest, as they say, is history. Only 30 Turkish ships returned to Constantinople in time for the delayed opening of the school year; and due to worsening weather and failure to deploy the proper biscuits in the proper advanced locations, The League did not press its advantage.

Despite the spectacular nature of the battle itself, let alone the unexpected one-sidedness of its result, nothing really changed in the balance of turf. Events suggest only that nobody got into more than minor fights for the next 20 years--at least in that part of the world.

Now, if the consequences were so slight, why is Lepanto still on our history examinations? Voltaire asked much the same question in the middle of the 18th century, and Fernand Braudel asked it in the middle of the 20th century, and they are both part of a long tradition of noted scholars asking the same question (which only demonstrates that historians can repeat themselves in more modes than tragedy and farce).

As an event, Lepanto is still on our examinations, I submit, because American higher education still has not figured out just what it is we want students to learn, and, more seriously, how to assess their learning. We can all join together, and effortlessly write the question for the GRE Area Test in History:

"Which of the following was not a consequence of the Battle of Lepanto?

(a) The Turkish fleet ceased to be a threat to the Christian states in the Mediterranean;
(b) Plans developed by The League to invade the Turkish Empire became more plausible;
(c) The Turkish Empire was so weakened as to put it on the road to collapse;
(d) Admiral Don John of Austria became a great hero among both sailors and emissaries to The League."

> As is the case with many similar questions we use in our tests--whether published and standardized or ad hoc classroom instruments--a student with sufficient ability to read subtle linguistic clues can discern the answer (c) without any prior knowledge of Lepanto or its century. Perhaps more telling, the form of this question (and so many others we use on examinations in which the item is the engine of assessment) calls for deductive reasoning, and treats its content as a fragment.

The example, and its brief analysis, raises a few issues I would like to cover in this concluding essay. Some of them have been mentioned in the course of other essays in this collection, while others have not. Some are "guidances" that would be found on any test-audit checklist that followed American Psychological Association standards for fair and ethical assessment practices. Others are technical concepts that ought to be given very concrete illustration.

But my over-arching point is that the ways in which we ask questions of students are often metaphors for the ways in which we teach and organize the stuff of learning. If one were to judge from items on GRE Subject Tests ranging from Biology to French, the history question I constructed is not an anomaly: the chances are high that we teach the Battle of Lepanto as a fragment. That it could be a minor blip of a vast screen of economic, meteorological, and cultural forces, and that it could be considered outside the category of causality--neither is assumed by our custom and usage in assessment, hence neither is likely in classroom presentation. As the following sections imply, we can do better.

Assessment of Second-Language and Other Special Populations

Millman, Dunbar, and Grandy have drawn the reader's attention to the criterion of construct validity, that is, the criterion by which we ensure that an examination designed to measure X in fact measures X and not Y. Put another way, we want to make sure that X is not so dependent on Y that we cannot measure it. There is no doubt that what Cummins (1980) called Cognitive Academic Language Proficiency in the language of instruction is essential to learning, particularly in higher education. The language of instruction is Standard American English. Generally speaking, the language of instruction is not literary English. If we construct questions, such as that for the Battle of Lepanto, where the ability to read literary English (to perceive that the style in which the correct answer is phrased differs considerably from the discursive presentation and neutral intensity of the alternatives) can lead a student to correct responses, then we are confounding the construct validity of the assessment. More than that, in a system of higher education that serves a nation of immigrants and the children of immigrants, those whose native or dominant household language is not English will be at a disadvantage on such assessments.

And with the increasing number of college students from households in which Oriental languages are spoken, we should be particularly watchful for wording of assessment tasks and items that relies too heavily on prepositions for comprehension.

Some academic administrators may respond that they have enough difficulty with the limited English proficiency of foreign graduate students serving as Teaching Assistants without worrying about undergraduates from second-language backgrounds. To set institutional standards for both populations, and hence to ensure congruence in both expression and comprehension in the classroom, it may be advisable to use the Foreign Service Institute's oral language proficiency interview and scales (discussed briefly in the essay on difficulty levels in general education assessments), with a minimum score of 4 for graduate instructors and 3+ for undergraduate students. The type of assessment we are wont to use for these populations (e.g. the TOEFL, Test of English as a Foreign Language) does not provide sufficient information concerning the criterion of academic language proficiency.

The writers and reviewers for this volume also take the position that special guidelines are necessary for the assessment of handicapped students (including those with learning disabilities). Particularly when the assessments involve decisions about individuals (e.g. placement or certification), speed should not be a criterion for people with disabilities that affect the quality of their performance under the constraints of time. In the practical terms of administering assessments, one must allow these students more time to complete the task.

One alternative for the assessment of students unable to take written examinations or to participate in assessments requiring any psychomotor skills in which they have a handicap, is the oral interview. The principal problem with this mode of assessment, of course, is its comparatively weak reliability. Its very virtues (e.g. the ability of the interlocutor to rephrase a question that the examinee does not seem to understand or to probe a response or to sample a variety of sources of the examinee's knowledge), become limitations in the lack of precision of such examining and in the inefficiency of the situation itself in terms of faculty and student time.

If oral assessment procedures are employed, rules should be set and strictly followed in order to be fair to the student and to produce as reliable results as possible under the circumstances of the method. As in other assessments, it is very important for faculty to write down the questions and determine the criteria for judging responses in advance, as well as to judge the student's performance with reference to those criteria, only. Providing examinees with sufficient time to respond, not probing excessively when the response to a given item is inadequate, and carefully but unobtrusively recording responses--all these steps will improve the technical adequacy and equity of the results.

Indeed, the point of this issue is that if we provide access to higher education, and if assessment is constitutive of higher education, we should also provide access to assessment.

Course-Level Assessment

I suspect higher education would not find itself under external pressure to develop assessment programs if faculty conducted technically sound and responsible assessments in their courses. Few faculty do, but we cannot blame them because fewer still were ever trained in matters of assessment. The question on the Battle of Lepanto is fairly well written as a test item, but neither isolated items nor restricted response tests (another, and more accurate, name for what we commonly call "objective" or "multiple-choice" tests) make for an appropriate assessment at the course-level.

Virtually every text on psychological measurement contains examples of poor classroom testing practices. These practices tend to exhibit problems of validity and/or sloppy construction of test items, ambiguous instructions, and the like. In higher education, the validity problem might be illustrated best by the test question that requires a student to draw on some relatively trivial piece of information from a textbook or on a minor point contained in a lecture. A test that assesses whether students have memorized every word of a textbook or have taken verbatim class notes is not a test of the subject.

Test items that repeat information from other test items or that provide grammatical clues to the correct answer (e.g. the stem of the question is in the singular and three out of four possible answers are phrased in the plural) or that offer other "cueing effects" (Newble, Baxter, and Elmslie, 1979) further confound the validity of the assessment. And essay examinations that offer statements or quotations and then instruct the student only to "discuss" can result in problems of reliability in grading, since criteria for the expected discussion would be difficult to set. (Chances are, they were never set in the first place.) The intent of an assessment task must be clear to the student, and not so generalized as to be fraught with ambiguity. Problem-solving tasks that are either too easy or too difficult or that contain superfluous details are also features of poorly designed classroom assessments. In all these cases, the relationship between the assessment and prior instruction is compromised.

The principles of assessment enunciated in this volume apply as much to assessment in the individual college classroom as they do across sections of the same course, departments, or programs. And because the results of classroom assessments are used to make judgments about individual students, particular care must be paid to questions of content validity, reliability, and bias. Since the writers of this volume take the position that assessment should be a mode of instruction as well as judgment, care should also be taken to ensure that classroom assessments yield information that is genuinely helpful to the student. That may mean narrative reports concerning student performance on a given instrument (even a final examination in a multi-section course). And if standardized multiple-choice tests are used, it definitely means more emphasis on item-type scores, sub-test scores, and criterion scores--all of which, by definition, are more descriptive (and hence instructionally useful) than nebulous global scores for a subject or intellectual skill.

From a technical perspective, classroom grading practices (like the process of supervisory ratings of employees) are subject to far more bias than any standardized test. What Stiggins and Bridgeford (1985) call "unstructured performance assessments" (in which neither objectives nor criteria for judgment are ever written down) are common practice, and are susceptible to behaviors and personality characteristics that have little to do with knowledge or skill. A course grade is a composite of judgments of discrete performances, and, as such, is analogous to an overall job rating. Kane (1986) distinguishes between iterated and noniterated job functions in terms of the frequency with which they occur during a given period of observation and assessment. A "noniterated function" is performed only once, like a final examination. "Iterated functions," such as lab set-ups and contributions to class discussions are performed more than once. Just as some students are "testwise," others are "gradewise," and engage in gaming behaviors with respect to iterated functions to influence faculty judgment of their academic work.

Given the variable and sometimes vulnerable judgment of faculty, these behaviors too often result in adversarial relationships that are detrimental to learning. It is for that reason, as Carmichael (1987) has suggested, that external examinations or examiners may sometimes be necessary for students and faculty to join together in proving their collective worth against a new kind of "adversary."

Given the minimal information provided to students by normal classroom grading practices, the alternative of formative classroom assessments may be worth consideration, though for the average college instructor they both violate inherited academic style and consume a great deal of time. By formative classroom assessments I mean carefully constructed tasks that provide information to students concerning what they have learned or can do with reference to summative performance criteria. The information, like that in the computer adaptive testing model that Grandy describes, is just that: information. The performance is not graded in the traditional sense. ("It does not count," is the way students would phrase it.) But its motivational value can be considerable if the student perceives the instructor as interested in his or her performance and learning, and the feedback is specific enough so that the student can perceive its relationship to performance (Ilgen, Fischer, and Taylor, 1979). What would count in such a system, though, would be the final course assessment.

Implicit in a number of essays in this volume is the strong suggestion that all multi-section courses develop and use formal test specifications for final examinations, and revise those specifications periodically to reflect changing faculty consensus on instructional objectives and emphases. This strategy enables a department to assess course and program effectiveness unobtrusively, to use its own students as a norm group, and to establish time-series data on the department's progress in teaching. In courses where external standardized examinations are available (e.g. the ACS exams for chemistry courses, CLEPS, PEPS, etc.), one can occasionally draw a matrix sample of students and compare results on the local exam to those on the nationally standardized exam as a check on the reliability of the former. In this process, of course, faculty should compare the test specifications and analyze the items on the standardized test in terms of their content distribution and level of difficulty. In that way, they will better understand what local test specifications indicate.

The Virtues and Limitations of Criterion-Referenced Inferences

Much of what we hear from faculty in terms of dissatisfaction with existing off-the-shelf tests that persist in asking fragmentary questions about the Battle of Lepanto, and much of the discussion in this volume that focuses on locally developed assessments refers to an alternative approach known as "criterion referenced testing" (or CRT, in the trade acronym). In many ways, a criterion-referenced assessment is not very different from a standardized published test--in both cases, a content domain must be described first by the test developers and tasks constructed to elicit student responses that will evidence mastery of the domain. The principal difference lies in the point of reference used when information on test results is analyzed and reported. For the CRT, it is the content domain as defined, and the "continuum of knowledge acquisition" within that domain (Glaser, 1963); for standardized published tests, it is the norm of the performance of other, similar students. In the CRT, the standards of content are absolute; in the norm-referenced test, the standards of performance are relative.

Unlike the intended use of some norm-referenced tests, that of criterion-referenced measurement is not predictive. Instead, the method was developed to describe the current status of an individual's proficiency with respect to a domain of knowledge irrespective of what that current status might say about future status. Given those reference points ("current status" and "domain of knowledge") the real value of the CRT approach is revealed in the construction of course or departmental examinations, for then faculty are forced to describe the content domain with great clarity and precision. It is thus more likely that the resulting assessment will reflect the emphases of local curricula, hence will evidence greater content validity in its context. By clarity and precision, though, I do not mean excessively detailed specifications. There is a point beyond which the specifications we set down in performance criteria negatively affect the reliability of judgment.

For purposes of program evaluation, it is far more helpful when faculty know that, for example, 72 percent of the graduating seniors in their department can perform X task at Y level of mastery as they have defined it; or that 38 percent of entering freshmen, but 88 percent of rising juniors can answer correctly 80 percent of the items on a vocabulary test based on analysis of materials typically read in a college's general education program; or that 44 percent of graduating history majors know something of the significance of 20 major events (including the Battle of Lepanto) in 16th-century Europe. These are criterion-referenced inferences.

Because of the way they are constructed, published tests do not exhibit these virtues. If the curriculum to which they refer is highly diverse, the content of the tests is very generalized, reflective of no particular local program. And since most published tests have been normed, many items that would be based on the portion of a curriculum common to many institutions (and hence that would produce high correct response rates) would never be included in the tests. The upshot, as Popham (1983) points out, is that program evaluations using standardized tests often wind up with "no significant difference" conclusions.

On the other hand, as Popham reports, when educators at the school or college level create their own criterion-referenced tests, the results have been "patently puerile" because "in addition to their pervasive psychometric shortcomings, . . . they demanded too little of students." It is a fraud to tell students or the general public that 95 percent of the students pass an assessment when the assessment itself does not embody what either students or the public expect from schooling. If we construct creative criterion-referenced assessments for college graduates that ask no more than what we would expect of high school graduates, then college students themselves should (and will) feel cheated.

One of the persistent problems with criterion-referenced methodologies in practice is that they ironically tell us more about how much a student knows or can do rather than how well he or she knows or does. Even if we specify a domain of intellectual skills (which rely on mastery of rules and manipulation of symbols) or verbal information (what we usually think of as facts stored for recall in verbal form), our task is incomplete without a specification of performance criteria (Gagne, 1974). If, for example, we wish to ascertain how well a student can analyze the rhetorical elements of a literary text, we should not set a performance criterion in such terms as "identifies and illustrates five devices contributing to the rhetorical effectiveness" of the text. Whether our performance criteria are explicit, that is unfortunately too often what we look for in a student's response to such a criterion-referenced task, and our judgment is really based on how much, not how well.

Performance, Judgment and Expertise

As the essays by Dunbar and Byham, in particular, have illustrated, the most difficult tasks in assessment lie in the design and execution of production measures--the systematic judgment of those human behaviors that directly embody the knowledge or skill that is the objective of instruction. In our common parlance, we refer to these tasks as performance assessments. In one sense, of course, all assessments--whether in the classroom or the workplace, whether in public service, entertainment, athletics, business, the professions, and such--involve performance and judgment (hence the title of this book). But in higher education we tend to oppose performance to paper-and-pencil tests, production measures to recognition measures, observable student behaviors to traces of knowledge concerning the Battle of Lepanto. As Dunbar notes in "States of the Art in the Science of Writing," the former are far more complex and psychometrically vulnerable than the latter.

There is great deal we do in assessment already that technically qualifies as performance. As Dunbar demonstrates in his essays, and as instructors of English composition know full well, writing itself is a performance. So is an assignment to develop and write down (i.e. provide observable evidence) a research design, or to construct a diagram of a process in engineering, or to compose a transition between two themes in a sonata. These are all "paper-and-pencil" operations bearing significant resemblance to restricted-response essays. In these cases, though, we pay far more attention to the product than the process, and where there is a product, an artifact, we can establish far more clear and public criteria for judgment than we are able to do for the process.

To be persuasive, judgments of performance must be as psychometrically precise as possible. But performance, considered as a process, is rarely subject to the scientific model of measurement. That is to say, performance, unlike a physical phenomenon such as an "objective" test: (a) does not possess a fixed value based on the average of measurements; and (b) exhibits variations around the average that are not the result of random measurement errors. The reasons for these differences should be obvious:

- o we cannot control the conditions of performance in the same way we can control a physical object or process;
- o performance is more likely to be influenced by motivation, which, as Graham points out in her essay, varies within individuals, let alone among them; and
- o we cannot provide to those who judge performance a fail-safe scale and code book of observations and their rules for each aspect of each criterion of the performance.

So we would like to be more global and impressionistic, the habit of academic judgment. Using job performance assessments, both Bernardin (1984) and Kane (1986) have demonstrated the viability of weighting schemes for components or "dimensions" of performance. The result is an excessive positivism, one with which academics would feel very uncomfortable, even though many will write, in a course syllabus, "the grade will be constructed from 20 percent for the final exam, 25 percent for the term paper" and so on.

The degrees of freedom one has in setting standards for performance differ by discipline, context, and task. In a performance in laboratory science, for example, one can call for a measure that is accurate within a given range or with reference to other quantitative benchmarks. In a musical or dramatic performance in which accuracy may again be a criterion, it is possible to set a range of mistakes or forgotten lines. Where speed, as well as accuracy, is a criterion (such as in a simulation of a hospital emergency room for students of medicine or nursing, or in simulations of international markets for students of finance), then obviously one should add to the criteria of performance the specification of a time frame.

The selection of a production measure is, in an adaptation of Millman's principles, driven by the purpose of the assessment. If we wish to know what students know about a subject, the production measure is neither as valid or efficient as a recognition measure. On the other hand, if we wish to know how competent students are in executing an activity, no recognition measure can replace a performance assessment. As one of the classic texts on the topic describes it (Fitzpatrick and Morrison, 1971), the performance task in an assessment is essentially a simulation of an actual situation. The higher the degree of realism in the situation, the more valid the assessment.

But in our fascination with performance assessments, we must recognize that the best we can do in an educational setting is to approximate behavior in a non-educational setting. Some tasks can be performed in both, but the conditions are decidedly different. As Stillman and Gillers (1986) observe of simulations in medical education, so-called "patient-management problems" may "evaluate skills of data gathering, hypothesis generation, and management," but are not comparable to performance in "the real clinical setting" (p. 397). The problem of evaluation in "the real clinical setting," of course, is that you can't standardize the patients. As Stillman and Gillers report, though, to the extent to which "patients" are taught to simulate an encounter and trained to evaluate intern performance on the basis of scales (for content and process) developed by expert physicians, there is a high degree of reliability in judgment. But the problem then shifts to a stronger toe, as other studies (e.g. Koran, 1975) have indicated that while experts can define the domains of performance in professional practice, they demonstrate significant disagreement concerning standards of performance within those domains.

There seems to be some fascination, too, with the assessment of "life experience"--the aposteriori version of what Byham describes in the assessment center. The major caveat concerning this type of performance assessment within the context of a higher education degree program lies in the problem of reflective understanding. Indeed, this problem lies at the core of all performance assessments. Observation of behavior in a specific context does not provide the certifying institution or faculty member with any assurance that the student can generalize from that behavior, that he or she can reflect on the context and describe its characteristics, or that he or she can abstract the principles of his or her own behavior and explain what kinds of behavior a different context might elicit. This same reflective judgment distinguishes the expert from the novice in, let us say, music performance. The more expert, the greater the student's ability to explain what he or she did in interpreting a classic concerto or in improvising on a jazz theme. Unless criteria for this reflective understanding are added to performance assessments, the assessments will not be wholly appropriate for the objectives of collegiate education.

Besides, if one of our principal objectives in improving assessment practices in higher education is to develop the capacity for self-assessment in students, reflective judgment in performance assessments should be required. The greater the range of concepts a student can use to judge his or her own performance, and the more reliably the student can apply those concepts, the more evidence we have that the student is moving across the spectrum from novice to expert.

Indeed, one of the implicit themes of this collection of essays (a theme that comes close to the surface in the pieces on assessment in the major, difficulty levels, and computer-based testing) is that the key to understanding the assessment of performance is the concept of expertise. Chi and Glaser (1980) have effectively presented this concept as a four-step challenge to psychologists, though the challenge can be offered equally to any disciplinary faculty involved in assessment:

- Identify what is involved in "high-level competence and expert performance;"
- Observe how people reach that level through other stages;
- Describe the cognitive and allied processes at each stage.

The fourth step is to describe each stage as a transformation from one quality of performance to another, that is: from variable performance to consistent performance; from consistent but fragmentary performance to strategic performance; from strategic but literal performance to symbolic performance. At that point, de facto benchmarks have been established, progress can be measured and learning facilitated. At that point--and with expertise as a reference--we will ask fragmentary questions about the Battle of Lepanto only to novices, if at all.

Student Participation in Assessment

The question of motivating students to participate in course and program assessments--when they have no personal, immediate stake in the outcome--continues to bedevil assessment planners and measurement specialists. The issue has been raised in a number of papers in this volume, with one author recommending paying the students and at least two of our reviewers vehemently objecting to that strategy. How much do you pay a student in exchange for what? Will $20 produce someone who will show up for the exam and perform to the maximum of his or her ability? Will $50 do it? We have not been able to answer these questions.

There are at least three alternative strategies, two of which introduce personal stake, hence greater motivation. The first of these would be for a college to require participation in one or two program assessments as a condition of graduation. But mere participation does not guarantee maximum effort. Alternatively, if one included special program assessment questions in the final examinations for individual courses--in the same manner as experimental questions are included in the different forms of standardized tests--and analyzed performance on those questions independently of performance on the whole examination, one is assured of a modicum of motivation on the part of students.

The writers of this volume all believe that student participation in assessment design and administration is essential to (though does not guarantee) assessment program success. There is no personal stake except for the interested students who volunteer to join faculty in these tasks; and such students will most likely participate (and speak with some authority) only in the context of assessment in their majors. Nonetheless, the symbolism of the participation of the few may result in positive effects on the attitudes of other students toward the assessment.

Conclusion: Assessment as Metaphor

The five issues briefly discussed in this concluding essay are not the only ones warranting consideration, but in light of the other essays in this volume, they deserved to be underscored. The language of assessment is another such issue, and no doubt deserves special treatment in a volume less concerned with practical advice and guidance. Here, though, it is worth noting that the words and phrases we use to talk about formal cognitive operations, testing, performance, and such have acquired emotive meanings in the contexts of heated exchanges between enthusiasts and adversaries. Anne Anastasi (1980) called them "excess meanings," and pointed out that they lead to misuses and misinterpretations and, ultimately, "disenchantment."

If assessment, as a process, involves performance, judgment, and learning, then higher education cannot afford such disenchantment. The writers of this book note that concern with shaping assessment as a learning in itself, long regarded as the domain of industry, the military, and experimental colleges, has now moved onto the center stage of American higher education. But we are still using the term "assessment" as a metaphor for external accountability pressures and blind demands to produce quantitative indices of institutional worth. The writers whose work appears here advocate a more comprehensive usage, an expansion of the metaphor to include learning. We urge those who understand this expanded metaphor to be generous with their knowledge and advice, but also to be respectful of the ways in which assessment already plays itself out in academic cultures, particularly in large, complex institutions.

At the same time, it would be foolish to deny the role of assessment in the society and economy outside colleges and universities, foolish to deny the core of our metaphorical usage. Judgments of quality performance and effectiveness will continue to be passed on individuals and organizations by an armada of licensing authorities, accreditation bodies, funding agencies, employers. Those who lead our colleges, community colleges, and universities can contribute to the current movement in ways that improve the quality of those judgments, or wait for the armada to find them in the Gulf of Lepanto. The rest, as they say, will be history.

References

Berk, R.A. Performance Assessment: Methods and Applications. Baltimore: Johns Hopkins University Press, 1986.

Bernardin, J. Performance Appraisal: Assessing Human Behavior at Work. Boston: Kent-Wadsworth, 1984.

Carmichael, J.W. "Standards With Sympathy." In Anderson, S.B. and Coburn, L.V. (eds.), Academic Testing and the Consumer. San Francisco: Jossey-Bass, 1982, pp. 39-44.

Chi, M.T.H. and Glaser, R. "The Measurement of Expertise: Analysis of the Development of Knowledge and Skill as a Basis for Assessing Achievement." In Baker, E.L. and Quellmalz, E.S. (eds.), Educational Testing and Evaluation: Design, Analysis and Policy. Beverly Hills: Sage Pubs., 1980, pp. 37-47.

Cummins, J. "The Cross-Lingual Dimensions of Language Proficiency." TESOL Quarterly, vol. 4 (1980), pp. 175-187.

Fitzpatrick, R. and Morrison, E.J., "Performance and Product Evaluation." In Robert L. Thorndike (ed.), Educational Measurement (2nd edition). Washington, D.C.: American Council on Education, 1971, pp. 237-270.

Gagne, R.M. "Task Analysis--Its Relation to Content Analysis." Educational Psychologist, vol. 11 (1974), pp. 11-18.

Glaser, R. "Instructional Technology and the Measurement of Learning Outcomes." American Psychologist, vol. 18 (1963), pp. 519-521.

Ilgen, D.R., Fischer, C.D. and Taylor, M.S. "Motivational Consequences of Individual Feedback on Behavior in Organizations." Journal of Applied Psychology, vol. 64 (1979), pp. 349-371.

Kane, J. "Performance Distribution Assessment," in Berk (1986), pp. 237-273.

Koran, L.M., "The Reliability of Clinical Data, Methods, and Judgment." New England Journal of Medicine, vol. 293 (1975), pp. 642-646 and 695-701.

Newble, D.I., Baxter, A. and Elmslie, R.C. "A Comparison of Multiple-Choice Tests and Free-Response Tests in Examinations of Clinical Competence." Medical Education, vol. 13 (1979), pp. 263-268.

Popham, W.J. "Measurement as an Instructional Catalyst." In Ruth B. Ekstrom (ed.), Measurement, Technology, and Individuality in Education. San Francisco: Jossey-Bass, 1983, pp. 19-30.

Stiggins, R.J. and Bridgeford, N.J., "The Ecology of Classroom Assessment." Journal of Educational Measurement, vol. 22 (1985), pp. 271-286.

Stillman, P.L. and Gellers, M.A., "Clinical Performance Evaluation in Medicine and Law," in Berk (1986), pp. 393-445.

Wigdor, A. and Garner, W.R. (eds.) Ability Testing: Uses, Consequences, and Controversies. 2 vols. Washington, D.C.: National Academy Press, 1982. Volume I: Report of the Committee [on Ability Testing].

Appendix A:

An Annotated Bibliography on the Assessment of Student Educational Outcomes

by Gary Pike

The 1970s and 1980s have witnessed a dramatic growth in the literature on the assessment of student educational outcomes. Because of the variety of information available, this bibliography is not intended to be exhaustive. Instead, this review is provided as a starting point for the study of assessment. For convenience, the literature on assessment is organized around four themes: the basic principles underlying assessment programs, the identification of educational outcomes, the measurement of these outcomes and the analysis of outcomes data. Where an ED number is indicated (in brackets), the document cited is available through the ERIC Document Reproduction Service, 3900 Willer Ave., Alexandria, Va. 22304.

Principles Underlying Assessment Programs

Adelman, C., "To Imagine an Adverb: Concluding Notes to Adversaries and Enthusiasts," in Adelman, C. (ed.), Assessment in American Higher Education: Issues and Contexts. Washington, D.C.: U.S. Government Printing Office, 1986, pp. 73-82. Assessment (evaluation) is rooted in the nature of language, and because evaluation inheres in language, assessment shapes and is shaped by social and economic institutions. Based on this perspective, the author identifies several important measurement, organizational and policy concerns related to the growing interest in assessment by institutions of higher education.

Baker, E.L., "Critical Validity Issues in the Methodology of Higher Education Assessment," Assessing the Outcomes of Higher Education: Proceedings of the 1986 ETS Invitational Conference. Princeton: ETS, 1987, pp. 39-46. Baker examines the growing interest in assessment, arguing that assessment programs designed to measure effectiveness criteria established by State governments and regional accrediting associations often do not represent valid means of evaluating educational quality. The author contends that students' classroom experiences represent the best indicators of quality. As a result, the author recommends that assessment programs focus on outcomes directly related to classroom experiences.

Bergquist, W.H. and Armstrong, J.L. Planning Effectively for Educational Quality: An Outcomes-Based Approach for Colleges Committed to Excellence. San Francisco: Jossey-Bass, 1986. The authors provide a model for improving educational quality based on an examination of institutional mission. They advocate that institutions examine their missions, develop pilot programs designed to assist in accomplishing those missions, assess the effectiveness of the pilot programs, and then implement large-scale programs. Of relevance to those interested in assessment, the authors stress the importance of incorporating outcomes data in the planning process.

Chandler, J.W., "The Why, What, and Who of Assessment: The College Perspective," Assessing the Outcomes of Higher Education: Proceedings of the 1986 ETS Invitational Conference. Princeton: ETS, 1987, pp. 11-18. According to Chandler, assessment should focus on programs, not individuals. Assessment also should reflect the unique characteristics of an institution. Tailoring an assessment program to an institution encourages faculty ownership of the assessment program. In addition, the author explains why assessment should not be equated with testing.

Cross, K.P., "Using Assessment to Improve Instruction," Assessing the Outcomes of Higher Education: Proceedings of the 1986 ETS Invitational Conference. Princeton: ETS, 1987, pp. 63-70. The author argues that evaluations of classroom teaching should be an integral part of an assessment program. The author notes that one of the best ways to overcome faculty resistance to an assessment program is to provide the faculty with the tools to assess student learning and satisfaction.

Enthoven, A.C., "Measures of the Outputs of Education: Some Practical Suggestions for Their Development and Use," in Lawrence, B., Weathersby, G., and Patterson, V.W. (eds.), Outputs of Higher Education: Their Identification, Measurement, and Evaluation. Boulder, CO: Western Interstate Commission for Higher Education, 1970, pp. 51-60. [ED#043-296] The author makes three recommendations about the assessment of educational outcomes: first, assessment should be coupled with financial incentives; second, external evaluations (rather than course examinations) should be used in the assessment program; and third, assessment activities should be conducted by a central office of program analysis and review.

Ewell, P.T., The Self-Regarding Institution: Information for Excellence. Boulder, CO: National Center for Higher Education Management Systems, 1984. This volume focuses on the rationale underlying the assessment of student outcomes. The author begins by identifying four dimensions of student outcomes: "Knowledge Outcomes," "Skills Outcomes," "Attitude and Value Outcomes," and "Relationships with Society and with Particular Constituencies." The author also provides examples of how institutions have utilized outcomes data in their planning processes to improve education programs.

Heywood, J. *Assessment in Higher Education*. New York: John Wiley, 1977. The author provides a general overview of assessment, focusing on two critical aspects of educational improvement: the specification of objectives and the measurement of the extent to which these objectives are being met. The author concludes that educational improvement will occur only if assessment is made an integral part of a process of curriculum development and evaluation. The argument is based primarily on assessment practices in the United Kingdom, but the principles are universally applicable. A second edition is scheduled for publication in 1988.

Loacker, G., Cromwell, L., and O'Brien, K., "Assessment in Higher Education: To Serve the Learner," in Adelman, C. (ed.), *Assessment in American Higher Education: Issues and Contexts*. Washington, D.C.: U.S. Government Printing Office, 1986, pp. 47-63. Loacker *et al* assert that the ultimate goal of an assessment program should be to promote student learning and development. The authors thus view assessment as a multitrait and multimethod technique, and stress the need to develop evaluation activities outside the traditional student-faculty process.

Manning, T.E., "The Why, What, and Who of Assessment: The Accrediting Association Perspective," in *Assessing the Outcomes of Higher Education: Proceedings of the 1986 ETS Invitational Conference*. Princeton: ETS, 1987, pp. 31-38. Manning examines assessment from the perspective of the regional accrediting associations, noting that these associations have been advocating the use of assessment to improve institutional quality for several years.

Identification of Educational Outcomes

Before developing an assessment program, institutions must identify the outcomes to be assessed. In an effort to bring some coherence to this undertaking, several scholars have developed typologies of educational outcomes. While these typologies differ in many important respects, they all assume that student outcomes are multidimensional. The common outcomes described in these typologies can be grouped into four categories: cognitive outcomes (both knowledge and skills); affective outcomes (such as self-concept and moral development); attitudinal outcomes (including involvement and satisfaction); and outcomes expressed in terms of longer-term economic and social status (and, sometimes, participation in cultural, community and political life).

Alexander, J.M. and Stark, J.S., *Focusing on Student Academic Outcomes: A Working Paper*. Ann Arbor, MI: National Center for Research to Improve Postsecondary Teaching and Learning, 1987. These authors provide an overview of three typologies of student educational outcomes (Astin, Panos, and Creager). In addition, they provide brief descriptions of instruments designed to assess student outcomes in three areas: academic-cognitive, academic-motivational, and academic-behavioral.

Astin, A.W., "Measuring Student Outputs in Higher Education," in Lawrence, B., Weathersby, G., and Patterson, V.W. (eds.), *Outputs of Higher Education: Their Identification, Measurement, and Evaluation*. Boulder, CO: Western Interstate Commission for Higher Education, 1970, pp.75-84. [ED#043-296] Astin discusses the measurement and analysis of educational outputs from a modeling perspective, presenting a classic model of the educational process consisting of three components: student inputs, the college environment, and student outputs. In addition, the author stresses the importance of conducting multitrait/multimethod research over time.

Astin, A.W., "The Methodology of Research on College Impact, Part One." *Sociology of Education*, vol. 43 (1970), pp. 223-254. Arguing that research on student development should consist of multi-institutional longitudinal studies, Astin identifies several research designs and statistical procedures that are appropriate for assessing student educational outcomes. Astin also discusses technical issues related to detecting interaction effects and controlling for the effects of measurement error.

Astin, A.W., "Measurement and Determinants of the Outputs of Higher Education," in Solmon, L. and Taubman, P. (eds.), *Does College Matter? Some Evidence of the Impacts of Higher Education*. New York: Academic Press, 1973, pp. 107-127. In this article, Astin discusses the relationships between types of outcome, data, and time. The first two dimensions form a taxonomy consisting of cognitive-psychological, cognitive-behavioral, affective-psychological, and affective-behavioral outcomes.

Bloom, B.S. (ed.) *Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain*. New York: David McKay, 1956. This classic work details a hierarchy of educational objectives, ranging from lower-order outcomes, such as knowledge recall, to higher-order outcomes, such as synthesis and evaluation. Examples of measurement techniques for evaluating the attainment of each level in the hierarchy are also provided.

Brown, D.G., "A Scheme for Measuring the Output of Higher Education," in Lawrence, B., Weathersby, G., and Patterson, V.W. (eds.) *Outputs of Higher Education: Their Identification, Measurement, and Evaluation*. Boulder, CO: Western Interstate Commission for Higher Education, 1970, pp. 27-40. [ED#043-296] Brown examines the outputs of higher education from a measurement perspective, identifying five categories of educational outcomes and six characteristics of effective measurement. Based on his categories, several specific measures are identified. He also presents a simple model that can be used to assess educational outcomes.

College Outcomes Evaluation Program, New Jersey Department of Higher Education. "Final Report of the Student Learning Outcomes Subcommittee." Trenton: Author, 1987. In its report, the subcommittee examines the purpose of statewide assessment in New Jersey and identifies the types of student outcomes to be assessed. These outcomes include the general intellectual skills needed to analyze and utilize new information, the skills needed to understand and use different modes of inquiry, and the abilities necessary to appreciate various "continuities in the human experience."

Korn, H.A. Psychological Models of the Impact of College on Students. Ann Arbor, MI: National Center for Research to Improve Postsecondary Teaching and Learning, 1987. Korn describes five perspectives on the relationship between college experiences and student educational outcomes, and discusses the implications of recent advances in personality theory for the assessment of student outcomes. Korn also suggests several ways in which the models can be used to evaluate the impact of college on students.

Lenning, O.T. Previous Attempts to Structure Educational Outcomes and Outcome-Related Concepts: A Compilation and Review of the Literature. Boulder: National Center for Higher Education Management Systems, 1977. This report provides a taxonomy of educational outcomes based on two literature reviews. Impacts of higher education on individuals include intellectual development, emotional/cultural/social development, and physical development. In addition, the author includes potential impacts of higher education on society.

Pace, C.R., "Perspectives and Problems in Student Outcomes Research," in Ewell, P.T. (ed.), Assessing Educational Outcomes. New Directions for Institutional Research No. 47. San Francisco: Jossey-Bass, 1985, pp. 7-18. Pace presents a general overview of basic assessment techniques and instruments, identifying four categories of outcomes, as well as instruments designed to measure these outcomes. Given the variety of outcomes and instruments that may be used in an assessment program, the author stresses the importance of selecting outcomes consistent with the institution's mission and goals.

Pascarella, E.T., "College Environmental Influences on Learning and Cognitive Development: A Critical Review and Synthesis," in Smart, J.C. (ed.), Higher Education: Handbook of Theory and Research. New York Agathon Press, 1985, pp. 1-62. Pascarella presents a comprehensive synthesis of research on the factors influencing students' cognitive development during their college careers. He defines two categories of cognitive outcomes (knowledge and skills), discusses research relating to each, and identifies several instruments that have been used to measure various educational outcomes.

Measurement of Educational Outcomes

Development of Measures

Once relevant student outcomes have been identified, some method of measurement must be selected or developed for each outcome. Assessment programs generally have relied on two types of measures: surveys and tests. Several scholars have emphasized the importance of basing test and survey development on empirical research. For surveys, empirical research can be used to identify variables of interest and pilot tests can evaluate item quality. Scholars also have argued that test domains should be derived empirically and item analysis should be used to evaluate item quality.

Dillman, D.A. Mail and Telephone Surveys: The Total Design Method. New York: John Wiley, 1978. Dillman presents an overview of survey research methodology. Specific topics addressed include question writing and formatting, sampling, questionnaire administration, data analysis, and reporting of results. Of particular interest to assessment practitioners, is Dillman's approach to issues of development and administration from the perspective of maximizing response rates.

Dumont, R.G. and Troelstrup, R.L., "Exploring Relationships Between Objective and Subjective Measures of Instructional Outcomes." Research in Higher Education, vol. 12 (1980), pp. 37-51. This article reports research designed to identify the relationship between test scores and self-reports of learning. The authors found that the two indicators evidence moderate positive correlations, and conclude that self-reports are valid measures of learning.

Ebel, R.L., "Content Standard Test Scores," Educational and Psychological Measurement, vol. 22, (1962), pp. 15-25. This author recommends that test scores be interpreted as content standard scores, indicating a student's level of mastery of a given content area. Ebel argues that content scores should be used to supplement normative scores, and provides an extended example of the derivation of content standard scores using the Preliminary Scholastic Aptitude Test (PSAT).

Frederiksen, N. and Ward, W.C. Development of Measures for the Study of Creativity. GRE Research Report GREB 72-2P. Princeton:

Educational Testing Service, 1975. In the research described in this report, four tests of scientific creativity were developed: formulating hypotheses, evaluating proposals, solving methodological problems, and measuring constructs. Results from a universe of 4,000 students applying to graduate schools indicated that the measures evidence acceptable levels of reliability. In addition, scores on each of the four measures were found to be independent of scores on aptitude and achievement tests.

Gronlund, N.E. *Constructing Achievement Tests*. 3rd edition. Englewood Cliffs, NJ: Prentice-Hall, 1983. This short book provides a basic introduction to the construction of achievement tests. The author addresses all phases of test preparation and evaluation, and discusses issues related to the construction and scoring of both objective and essay tests.

Grosof, M.S. and Sardy, H. "Procedure: Measurement, Instrumentation, and Data Collection," in *A Research Primer for the Social and Behavioral Sciences*. Orlando, FL: Academic Press, 1985, pp. 133-168. These authors provide an overview of several measurement techniques, including surveys. They identify the various types of questions used in survey research and describe several approaches to scaling. They also provide several basic recommendations regarding question wording and discuss approaches to evaluating questionnaire reliability and validity.

Hambleton, R.K., "Determining Test Length," in Berk, R. A. (ed.), *A Guide to Criterion-Referenced Test Construction*. Baltimore: Johns Hopkins University Press, 1984, pp. 144-168. Hambleton notes that test length has important implications for the reliability and validity of criterion-referenced tests. Five different methods of determining test length are described, and factors influencing the selection of one of these methods are identified.

Marshall, J.C. and Hales, L.W. *Essentials of Testing*. Reading, MA: Addison-Wesley, 1972. Marshall and Hales provide a nontechnical discussion of a variety of approaches to test construction. In addition to identifying several principles of educational measurement, the authors detail the strengths and weaknesses of essay tests, completion tests, multiple-choice tests, and true-false tests.

Martuza, V.R. *Applying Norm-Referenced and Criterion-Referenced Measurement in Education*. Boston: Allyn and Bacon, 1977. Martuza describes the use of norm-referenced and criterion-referenced tests in educational research. Regarding norm-referenced tests, Martuza explains the importance of selecting appropriate norm groups, provides criteria for evaluating norms, and provides a step-by-step guide for test construction. Martuza also suggests several approaches to constructing criterion-referenced exams, including linguistic transformation, item-form/item-frame, amplified objectives, and facet design.

Mehrens, W.A. and Ebel, R.L. "Some Comments on Criterion-Referenced and Norm Referenced Achievement Tests." NCME Measurement in Education, vol. 10, (1979), pp. 1-8. [ED#182-324] The authors discuss two approaches to achievement testing: norm-referenced and criterion-referenced tests. In addition to defining these two types of tests, the authors conclude that norm-referenced tests are most appropriate for evaluating curriculum, while criterion-referenced exams are most appropriate for evaluating students' mastery levels.

Milton, O. and Eison, J.A. Textbook Tests: Guidelines for Item Writing. New York: Harper and Row, 1983. This is a basic introduction to writing test items. The authors underscore the importance of well-designed tests and offer several practical suggestions concerning item writing. They also include a series of exercises that allow the reader to identify the weaknesses of test questions.

Popham, W.J. "Specifying the Domain of Content or Behaviors," in Berk, R.A. (ed.), A Guide to Criterion-Referenced Test Construction. Baltimore: Johns Hopkins University Press, 1984, pp. 49-77. Popham addresses the issue of how to specify the areas of content and/or behavior to be covered in a test, stressing the importance of explicit test specification and congruent test item development. The author also makes several practical suggestions regarding the specification process that have implications for subsequent steps in the test development process.

Roid, G.H. "Generating the Test Items," in Berk, R.A. (ed.), A Guide to Criterion-Referenced Test Construction. Baltimore: Johns Hopkins University Press, 1984, pp. 49-77. Roid reviews several item-writing techniques and argues that the quality of the items generated in the test construction process can be enhanced if the items are based on empirical research. Four steps in the empirically derived item-writing process are identified.

Macro-Evaluation of Measures

Macro-evaluation of student outcomes measures is concerned with the reliability and validity of these measures. There are many approaches to evaluating the reliability of outcomes, ranging from classical correlational techniques to techniques that assess the internal consistency of measures based on generalizability theory. Because assessment efforts frequently have multiple purposes, multiple approaches to evaluating instrument reliability frequently are necessary.

The second major criterion for evaluating assessment instruments is validity. Instruments can be evaluated in terms of their content validity, criterion-related validity, and their construct validity. As with reliability, the type of validity evaluated may change depending on the purpose of the assessment program.

Anastasi, A. Psychological Testing. 4th edition. New York: Macmillan, 1976. This book is a basic reference work on the development, use, and evaluation of psychological tests. Topics addressed include ethical issues in the use of psychological tests, evaluation of instrument reliability and validity, and item analysis. In addition, the author identifies and analyzes several different types of tests, ranging from educational (achievement) tests to personality measures.

Berk, R.A. (ed.) A Guide to Criterion-Referenced Test Construction. Baltimore: Johns Hopkins University Press, 1984. This book contains essays that provide a technical discussion of the construction and evaluation of criterion-referenced tests. Essays on the evaluation of tests address issues of reliability and validity, noting that the decision to utilize a specific approach must be guided by the intended uses of the test data. In addition, essays on evaluating the reliability of cut-off scores and categorizations based on cut-off scores are included.

Cronbach, L.J. "Test Validation," in Thorndike, R.L. (ed.), Educational Measurement. 2nd edition. Washington, D.C.: American Council on Education, 1971, pp. 443-507. Cronbach's essay is a touchstone for understanding test validity. The author explains the goals of validation procedures and examines several types of validity: content validity, educational importance, construct validity, validity for selection, and validity for placement.

Cronbach, L.J. and Meehl, P.E. "Construct Validity in Psychological Tests." Psychological Bulletin, vol. 52, (1955), pp. 281-302. These authors examine procedures for validating psychological tests, focusing on construct validity. They indicate when construct validation of tests is appropriate and examine the assumptions underlying construct validity.

Gardner, E. "Some Aspects of the Use and Misuse of Standardized Aptitude and Achievement Tests," in Widgor, A.K. and Garner, W.R. (eds.), Ability Testing: Uses, Consequences, and Controversies: Part II. Washington, D.C.: National Academy Press, 1982, pp. 315-332. Gardner identifies six categories of misuse associated with an unquestioning reliance on standardized tests: acceptance of the test title for what the test measures; ignoring the error of measurement in test scores; use of a single test score for decision making; lack of understanding of test score reporting; attributing cause of behavior measured to the test; and test bias.

Linn, R. "Ability Testing: Individual Differences, Prediction, and Differential Prediction," in Widgor, A.K. and Garner, W.R. (eds.), Ability Testing: Uses, Consequences, and Controversies: Part II. Washington, D.C.: National Academy Press, 1982, pp. 335-388. This essay examines the use of standardized tests to assess individual differences. The author addresses issues related to criterion and predictive validity for educational and occupational performance, and the effects of socioeconomic and racial/ethnic differences.

Mehrens, W.A. and Lehmann, I.J. Using Standardized Tests in Education. 4th edition. New York: Longmann, 1987. Mehrens and Lehmann provide a general overview of measurement and evaluation in education. The chapter on reliability discusses approaches to estimating reliability based on correlational and generalizability theories. The chapter on validity identifies several different types of validity and presents methods for their estimation.

Stanley, J.C. "Reliability," in Thorndike, R.L. (ed.), Educational Measurement. 2nd edition. Washington, D.C.: American Council on Education, 1971, pp. 356-442. This basic reference on estimating reliability in educational measurement examines its topic in light of research on individual variation, and identifies sources of variation in test scores. The author also presents procedures for estimating reliability using classical correlational techniques and generalizability theory and discusses methods of estimating the reliability of change scores.

Wigdor, A.K. and Garner, W.R. (eds.) Ability Testing: Uses, Consequences, and Controversies: Part I. Washington, D.C.: National Academy Press, 1982. Part I of this work is the report of the Committee on Ability Testing of the Assembly of Behavioral and Social Sciences, National Research Council. The report provides an overview of ability testing (including the controversies associated with ability testing), identifies the uses of ability tests, and recommends a series of actions for the evaluation and improvement of ability tests.

Micro Evaluation of Measures

Micro-evaluation of outcomes measures is concerned with the analysis of individual questions (items). Several procedures are available to analyze test items, ranging from relatively simple item analysis procedures to mathematically sophisticated procedures based on Item Response Theory (IRT). Approaches based on IRT offer significant advantages (e.g., item difficulty estimates that vary according to the ability level of the student). IRT approaches also have important applications in detecting test item bias, equating test scores, and in developing tailored and computer-adaptive tests.

Berk, R.A. "Conducting the Item Analysis," in Berk, R.A. (ed.), *A Guide to Criterion-Referenced Test Construction*. Baltimore: Johns Hopkins University Press, 1984, pp. 97-143. Berk presents a technical discussion of the procedures that should be used to determine if individual test items function as they were intended. He emphasizes that both expert judgment and statistical techniques should be used to evaluate test items. In addition to providing a discussion of specific judgmental and statistical tests, he identifies step-by-step procedures for item analysis.

Diederick, P. *Short-Cut Statistics for Teacher-Made Tests*. Princeton: ETS, 1973. The author presents an introduction to the analysis of item quality for the less sophisticated mathematician. Topics addressed in the text include reliability, measurement error, and item analysis.

Hambleton, R.K. and Cook, L.L. "Latent Trait Models and Their Use in the Analysis of Educational Test Data." *Journal of Educational Measurement*, vol. 14, (1977), pp. 75-96. This article represents a general introduction to the use of latent trait (item response) models in education research. The authors begin by identifying the fundamental principles underlying latent trait theory, identify several common latent trait models, and suggest several applications for these models.

Hambleton, R.K. and Swaminathan, H. *Item Response Theory: Principles and Applications*. Boston: Kluwer-Nijhoff, 1985. The authors provide a basic reference work on item response theory. Topics addressed include ability scales, model fitting, and practical applications of item response theory.

Lord, F.M. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1980. In this technical discussion of item response theory. Lord identifies several applications of IRT, including tailored testing, ability testing, studies of item bias, and estimation of true-score distributions.

Office for Minority Education. *An Approach for Identifying and Minimizing Bias in Standardized Tests: A Set of Guidelines*. Princeton: ETS, 1980. This report explains the issues related to bias in testing, and presents a series of guidelines for eliminating item bias in test construction and evaluating existing tests to detect biased items.

Assessment of Writing/Using Essay Examinations

Breland, H.M., Camp, R., Jones, R.J., Morris, M.M., and Rock, D.A. *Assessing Writing Skill*. Research Monograph No. 11.

New York: College Entrance Examination Board, 1987. These authors describe a study designed to assess writing skill at six colleges and universities. Results indicated that the unreliability of essay scoring could be alleviated by relying on multiple essays or by combining objective and essay tests. The authors also demonstrate the use of a variety of data analysis techniques. Both essay and objective tests were found to be about equal in their predictive validity. The authors conclude that multi-method assessment techniques offer both theoretical and practical advantages over other approaches.

Coffman, W.E. "Essay Examinations," in Thorndike, R.L. (ed.), Educational Measurement. 2nd edition. Washington, D.C.: American Council on Education, 1971, pp. 271-302. In this chapter, the author examines the advantages and limitations of essay tests as assessment tools, with specific attention to issues related to the reliability and validity. In addition, the author offers several suggestions for improving the use of essay exams.

Coffman, W.E. "On the Validity of Essay Tests of Achievement." Journal of Educational Measurement, vol. 3, (1966), pp. 151-156. This author reports research concerning methods of validating essay and objective tests. Traditionally, essay and objective tests have been correlated in order to demonstrate the predictive validity of objective tests. The author examines the predictive power of a sample of essay questions independent of objective measures.

Cooper, P.L. The Assessment of Writing Ability: A Review of Research. GRE Research Report GREB 82-15R. Princeton: ETS, 1984. The psychometric and practical issues related to the assessment of writing are the focus of this review. The author notes that although essay tests are considered to be more valid than multiple-choice tests, variability in subjects' scores may be influenced by a wide range of irrelevant factors. The author contends that when procedures to correct for threats to reliability and validity are employed, essay tests correlate very highly with multiple-choice tests.

Crocker, L. "Assessment of Writing Skills Through Essay Tests," in Bray, D. and Belcher, M. J. (eds.), Issues in Student Assessment. New Directions for Community Colleges, No. 59. San Francisco: Jossey-Bass, 1987. In discussing the use of essay tests in assessing basic writing skills, the author provides a rationale for using essay exams to assess writing abilities and identifies the steps required to develop a writing assessment program. These steps include: developing prompts (topics), developing scoring procedures, training raters, field testing, and administering the instruments. The author also examines issues related to the reliability and validity of essay exams.

Keeley, S.M., Browne, N.M., and Kreutzer, J.S. "A Comparison of Freshmen and Seniors on General and Specific Essay Tests of Critical Thinking." Research in Higher Education, vol. 17, (1982), pp. 139-154. These authors report research utilizing essay tests to evaluate the critical-thinking skills of freshmen and seniors. Results indicate that educational experiences produce significant gains in critical-thinking skills. An important finding for assessment practitioners was that significant differences in students' writing samples are related to the type of instructions (general or specific) provided for the assessment.

Steele, J.M. "The Assessment of Writing Proficiency via Qualitative Ratings of Writing Samples." Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, 1979. [ED#175-944] Steele examines several strategies for improving the reliability of raters' evaluations of writing samples. Research has indicated that increasing the number of writing samples per student to three significantly increases interrater reliability. However, using more than two raters does not improve reliability significantly.

Steele, J.M. "Trends and Patterns in Writing Assessment." Paper presented at the Annual Conference on the Assessment of Writing." San Francisco, 1985. [ED#268-146] The author describes the writing assessment portion of the College Outcome Measures Project (COMP) Composite Examination. He notes that the COMP exam, unlike many writing assessment instruments, focuses on writing in problem solving and critical thinking situations. Instead of providing a single holistic rating, the COMP writing assessment provides scores in three areas of writing proficiency.

White, E.M. Testing and Assessing Writing. San Francisco: Jossey-Bass, 1985. This book offers an overview of issues related to the assessment of writing. Included are discussions of holistic scoring, the use of proficiency tests, selection and/or development of writing tests, and the evaluation/scoring of writing assignments.

Nontraditional Outcomes Measures

During the last decade, there has been a marked increase in the use of nontraditional approaches to assess student educational outcomes. As a general rule, these approaches have been intended as supplements to existing measurement techniques. Reliance on multiple assessment methods has been shown to improve the validity of evaluations.

Most of the nontraditional measurement approaches have focused on the assessment of student performance through such techniques as assessment centers, simulations, and external evaluators. Exceptions have included computer-adaptive testing methods and the use of unobtrusive (nonreactive) measures to gather assessment data.

Berk, R.A. (ed.) Performance Assessment: Methods and Applications. Baltimore: Johns Hopkins University Press, 1986. This basic technical reference work on performance assessment includes essays covering a variety of performance assessment methods ranging from behavior rating scales to assessment center techniques. The authors also identify applications of performance assessment in business, medicine and the law, teaching, and the evaluation of communication skills.

Fong, B. The External Examiner Approach to Assessment. Washington, D.C.: AAHE Assessment Forum, 1987. This monograph provides an overview of the use of external examiners as an assessment tool. While both British and American experience is considered, special attention is paid to how American institutions are using external examiners to evaluate student mastery of content in courses and disciplines. The author also discusses issues of reliability and validity as they relate to the use of external examiners.

Hsu, T. and Sadock, S.F. Computer-Assisted Test Construction: The State of the Art. Princeton: ETS, 1985. [ED#272-515] These authors discuss both theory and applications of computers in developing and administering tests. They contend that adaptive testing is the one example of the successful use of computers to improve the quality of the assessment process.

Millman, J. "Individualizing Test Construction and Administration by Computer," in Berk, R.A. (ed.), A Guide to Criterion-Referenced Test Construction. Baltimore: Johns Hopkins University Press, 1984, pp. 78-96. Millman presents a technical review of the application of computers in test construction and administration. Specific topics include traditional attempts to individualize testing (equivalent forms of a test), item banking, and computer-adaptive testing. Millman notes that the proliferation of computer-adaptive tests has created a need for further research on the cost effectiveness of this approach. Millman concludes that assessment practitioners should be very cautious in utilizing computer-adaptive tests developed outside their own institutions.

Stillman, P.L. and Swanson, D.B. "Ensuring the Clinical Competence of Medical School Graduates Through Standardized Patients." Archives of Internal Medicine, vol. 147, (1987), pp. 1049-1052. These authors discuss the use of "Standardized Patients" to assess medical students' interviewing and physical examination skills. "Standardized Patients" are trained to function in multiple roles and to simulate a physician-patient encounter. Preliminary research suggests that this approach offers a realistic means of standardizing performance assessment for medical school graduates.

Terenzini, P.T. "The Case for Unobtrusive Measures," *Assessing the Outcomes of Higher Education: Proceedings of the 1986 ETS Invitational Conference*. Princeton: ETS, 1987, pp. 47-61. Terenzini argues that traditional data collection methods (tests, surveys, and interviews) should be supplemented by unobtrusive measurement techniques that can overcome the sources of measurement error present in other approaches, and are relatively inexpensive to administer. The author also presents a typology of unobtrusive measurement techniques that can be used to guide the selection of particular measures.

Urry, V.W. "Tailored Testing: A Successful Application of Latent Trait Theory." *Journal of Educational Measurement*, vol. 14, (1977), pp. 181-196. Urry describes the role of Item Response Theory in the development and administration of tailored (computer-adaptive) tests. In addition, he analyzes the computer-adaptive test used by the U.S. Civil Service Commission and identifies future uses for computer-adaptive ability tests.

Webb, E.J., Campbell, D.T., Schwartz, R.D., and Sechrest, L. *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally, 1966. In this classic short work, the authors present several reasons for supplementing traditional measurement techniques with unobtrusive measures, and identify several approaches to unobtrusive measurement. These measurement techniques include physical traces, archival data, simple observations, and contrived observation.

Webb, E. and Weick, K.E. "Unobtrusive Measures in Organization Theory: A Reminder," in Maanen, J.V. (ed.), *Qualitative Methodology*. Beverly Hills, CA: Sage Publications, 1983, pp. 209-224. In this chapter, the authors examine the use of unobtrusive measures in organizational research. Of particular interest to assessment practitioners, the authors identify six ways in which unobtrusive measurement can modify traditional data collection methods.

Value-Added Analysis of Outcomes Data

While the concept of the value added by a college education is compelling, several scholars have criticized the concept when value added is defined as a gain or difference score. Writers have suggested several alternatives to simple gain, including residual and base-free measures of gain. Still other scholars have suggested that repeated measures designs be used in value-added analyses.

Bereiter, C. "Some Persisting Dilemmas in the Measurement of Change," in Harris, C.W. (ed.), Problems in Measuring Change. Madison: University of Wisconsin Press, 1963, pp. 3-20. This introduction to the problems inherent in the use of change scores identifies and analyzes three dilemmas associated with their use: over-correction/under-correction; unreliability/ invalidity; and physicalism/subjectivism.

Cronbach, L.J. and Furby, L. "How We Should Measure 'Change'-- Or Should We?" Psychological Bulletin, vol. 74, (1974), pp. 68-80. See also "Errata." Ibid., p. 218. This technical discussion of methods for calculating change (difference) scores offers formulas for calculating "true" scores, "residual" scores, and "base-free" measures. Given the purposes for which change scores are used, the authors recommend a multivariate approach for evaluating change. Researchers interested in using the formulas presented in this article should carefully read the "Errata."

DuBois, P.H. "Correlational Analysis in Training Research," in DuBois, P.H. and Mayo, G.D. (eds.), Research Strategies for Evaluating Training. Chicago: Rand McNally, 1970. DuBois discusses the use of change scores in correlational research, explains the rationale underlying residual change scores, and presents a formula for calculating residual gain scores.

Gaito, J. and Wiley, D.E. "Univariate Analysis of Variance Procedures in the Measurement of Change," in Harris, C.W. (ed.), Problems in Measuring Change. Madison: University of Wisconsin Press, 1963, pp. 60-84. The authors present a basic description of the use of univariate analysis of variance to analyze repeated measures data. They begin by describing the assumptions underlying the univariate analysis of variance and then present a mathematical explanation of the univariate model. The authors also identify several procedures that may be used to minimize the effects of contaminating influences.

Horst, P. "Multivariate Models for Evaluating Change," in Harris, C.W. (ed.),Problems in Measuring Change. Madison: University of Wisconsin Press, 1963, pp. 104-121. Horst describes the theory underlying a general multivariate model for the evaluation of change. Horst examines the assumptions underlying a multivariate approach and provides a mathematical model for the multivariate analysis of change. This model relies on a multi-categorical matrix in which row vectors represent subjects and column vectors represent administrations of an instrument.

Lord, F.M. "Elementary Models for Measuring Change," in Harris, C.W. (ed.), Problems in Measuring Change. Madison: University of Wisconsin Press, 1963, pp. 21-38. Lord examines several problems inherent in the use of difference scores, including unreliability and regression effects. He notes that these problems can produce spurious relationships between gain scores and other variables;

and presents a method for calculating true gain scores.

McMillan, J.H., "Techniques for Evaluating Value-Added Data: Judging Validity, Improvement, and Causal Inferences." Paper presented at the annual meeting of the American Educational Research Association, Washington, 1987. The author identifies several limitations of value-added analyses and describes methods of overcoming these limitations. He suggests that researchers utilize appropriate research designs and statistical procedures when evaluating difference scores. One possible source of confirmatory data would be faculty judgments.

Pascarella, E.T., "Are Value-Added Analyses Valuable?" *Assessing the Outcomes of Higher Education: Proceedings of the 1986 ETS Invitational Conference*. Princeton: ETS, 1987, pp. 71-92. Pascarella presents a nontechnical discussion of the benefits and problems of relying on value-added data. He suggests several different methods of overcoming the problems associated with the use of difference scores and presents modeling techniques that can be used with value-added data.

Rogosa, D., Brandt, D., and Zimowski, M. "A Growth Curve Approach to the Measurement of Change." *Psychological Bulletin*, vol. 92, (1983), pp 726-748. These authors argue that the criticism of change scores as unreliable does not mean that they should be abandoned.

Tucker, L.R., Damarin, F., and Messick, S. "A Base-Free Measure of Change." *Psychometrika*, vol. 31, (1966), pp. 457-473. In this article, the authors identify and discuss problems with the calculation and use of simple gain scores. They recommend the use of a base-free measure of change, and provide formulas for calculating this measure.

Willett, J.B., "Questions and Answers in the Measurement of Change," in Rothkopf, E.R. (ed.), *Review of Research in Education*. volume 14. Washington, D.C.: American Educational Research Association, 1987. According to Willett, the measurement and analysis of growth (change) is central to evaluating educational effectiveness. Willett contends that the criticisms of growth measures that have been directed at two-wave (pre- and posttest) designs are overstated. Although Willett identifies instances in which simple difference scores can be reliable and valid, he recommends a multi-wave approach to measuring change.

Wolfle, L.M. "Applications of Causal Models in Higher Education," in Smart, J.C. (ed.), *Higher Education: Handbook of Theory and Research*. New York: Agathon Press, 1985. Wolfle examines the use of causal modeling as a research tool in the assessment of educational outcomes, explaining its assumptions and analyzing the concepts of causation and the decomposition of effects. The author also discusses the specification of recursive and nonrecursive models, and the use of causal models with latent variables.

Appendix B:

Review of Assessment Instruments

by Gary Pike

This review is designed to provide brief descriptions of the technical characteristics of many of the instruments mentioned in this book. For convenience, the descriptions are organized around six types of outcomes: general education, basic skills, cognitive development, learning in the discipline, values, and motivation. Within each outcome area, tests are listed in alphabetical order.

Assessment of General Education

Academic Profile

Publisher: ETS College and University Programs, Educational Testing Service, Princeton, NJ 08541-0001; Scales: Total Score, Humanities, Social Science, Natural Sciences, Reading, Writing, Critical Thinking, and Mathematics; Length: 48-144 items; Time: 1-3 hours.

The Academic Profile has been developed by ETS and the College Board to assess the effectiveness of general education programs. The Academic Profile is available in two forms: a one-hour exam providing group feedback, and a three-hour exam providing individual feedback. A panel of experts in the content fields supervised test construction, assisting with questions of content validity. Because ETS is making the Academic Profile available for pilot testing during the 1987-1988 academic year, further information about the reliability and validity of this test is not available at this time.

ETS College and University Programs. The Academic Profile. Princeton: ETS, 1981.

ACT Assessment Program

Publisher: American College Testing Program, P.O. Box 168, Iowa City, IA 52240; Scales: Composite Score, English Usage, Mathematics Usage, Social Studies Reading, Natural Science Reading; Length: 40-75 items/test; Time: 30-50 minutes/test.

The ACT Assessment Program was developed as a series of college entrance and placement examinations for high school graduates. Depending on the coefficients used, reliability estimates have ranged from .73 to .91. Research has found that the ACT Assessment is capable of predicting subsequent performance in college, including cumulative grade point average and performance in specific classes. However, research at Tennessee Technological University could not demonstrate a relationship between gains on the ACT Assessment exam and students' experiences in college, raising questions about the validity of the ACT Assessment exam as a measure of educational effectiveness.

American College Testing Program. *Assessing Students on the Way to College: Technical Report for the ACT Assessment Program*. Iowa City, IA: ACT, 1973.

American College Testing Program. *College Student Profiles: Norms for the ACT Assessment*. Iowa City, IA: ACT, 1987.

Dumont, R.G. and Troelstrup, R.L. "Measures and Predictors of Educational Growth with Four Years of College." *Research in Higher Education*, vol. 14, (1981), pp. 31-47.

Munday, L.A. "Correlations Between ACT and Other Predictors of Academic Success in College." *College and University*, vol. 44, (1968), pp. 67-76.

Richards, J.M., Jr., Holland, J.L., and Lutz, S.W. "Prediction of Student Accomplishment in College." *Journal of Educational Psychology*, vol. 58, (1967), pp. 343-355.

College Basic Academic Subjects Examination

Publisher: Center for Educational Assessment, University of Missouri-Columbia, 403 South Sixth Street, Columbia, MO 65211; *Scales*: English, Mathematics (2), Science, Social Studies, Reading, Reasoning, and Writing (optional); *Length*: approximately 40-120 items; *Time*: 1-3 hours.

The College Basic Academic Subjects Examination (College BASE) is a criterion-referenced achievement test that can be used to evaluate individuals or programs. One-and three-hour forms of the exam are available. Content validity of the College BASE was achieved by using expert reviewers during the test construction process. Because the exam is being pilot tested during the 1987-88 academic year, additional information on reliability and validity has not been made available.

Center for Educational Assessment. *College BASE*. Columbia, MO: University of Missouri-Columbia, 1987.

Collegiate Assessment of Academic Proficiency

Publisher: American College Testing Program. 2201 N. Dodge St., P.O. Box 168, Iowa City, Iowa 52243; *Scales*: Reading, Mathematics, Writing, and Critical Thinking. *Length*: 175 items for all four modules in pilot administration plus 2 prompts for writing sample; Time: 40 minutes for each module and 40 minutes for the writing sample.

The Collegiate Assessment of Academic Proficiency (CAAP) is a new standardized test intended to assist institutions in evaluating their general education programs by assessing those academic skills typically developed during the first two years of college. The CAAP is available in modules, and institutions may add questions to the exam, thereby tailoring the exam to their curriculum. Because the exam is being pilot-tested beginning in 1988, information on reliability and validity is not available.

American College Testing Program. Collegiate Assessment of Academic Proficiency: Test Specifications and Sample Items. Iowa City, IA: ACT, 1988.

CLEP Education Assessment Series

Publisher: The College Board. 45 Columbus Ave. New York, NY 10023-6917. Scales: English Composition, Mathematics; Length: 40-45 questions per scale; Time: 45 minutes per module.

The Education Assessment Series (EAS) consists of two tests intended to provide comprehensive, nationally-normed data in a relatively short administration time and at low cost. Because multiple forms of the exams will be available, institutions may administer them twice and calculate the "value added" by general education. The tests are being piloted in 1988, hence information concerning reliability and validity is not yet available.

The College Board. CLEP Introduces the Education Assessment Series. New York: Author, 1988.

CLEP General Education Examinations

Publisher: College Entrance Examination Board, 45 Columbus Ave. New York, NY 10023-6917; Tests: English Composition, Humanities, Mathematics, Natural Science, and Social Science/History; Length: 55-150 items/test; Time: 90 minutes/test.

The College-Level Examination Program (CLEP) General Examinations cover five content areas and were designed to provide college credit for non-college learning. Reliabilities for the five tests range from .91 to .94. Using panels of experts in the content fields, the CLEP test development process has achieved satisfactory levels of content validity. While research has linked CLEP scores to performance in introductory college courses, no studies have been conducted on the validity of the CLEP exams as program evaluation instruments.

College Entrance Examination Board. Technical Manual Overview. Princeton: ETS, 1984.

College Entrance Examination Board. Outcomes Assessment in Higher Education. Princeton: ETS, 1986.

College Outcome Measures Project

Publisher: ACT, P.O. Box 168, Iowa City, IA 52243; Scales: Total Score, Functioning within Social Institutions, Using Science and Technology, Using the Arts, Communicating, Solving Problems, Clarifying Values, Writing (CE), Speaking (CE), Reasoning and Communicating (CE); Length: 60-99 items; Time: 2.5-4.5 hours.

The College Outcome Measures Project (COMP) examination was designed to measure the knowledge and skills necessary for effective functioning in adult society. This exam is available in two forms: the Objective Test (OT), consisting of 60 multiple-choice items; and the Composite Examination (CE), containing the same multiple-choice questions and speaking/writing exercises. Estimates of reliability for the COMP sub-scales were satisfactory (ranging from .63 to .81) although research on its validity as an assessment instrument has produced mixed results. Studies by ACT have shown that COMP scores are related to general education coursework and student involvement; however, other research by colleges themselves has failed to find a link between COMP scores (or gains on the COMP) and effective academic programs.

Banta, T.W., Lambert, E.W., Pike, G.R., Schmidhammer, J.L. and Schneider, J.A., "Estimated Student Score Gain on the ACT COMP Exam: Valid Tool for Institutional Assessment?" Paper presented at the annual meeting of the American Educational Research Association, Washington, 1987. [ED#281-892]

Forrest, A. Increasing Student Competence and Persistence: The Best Case for General Education. Iowa City, IA: ACT National Center for the Advancement of Educational Practices, 1982.

Forrest, A. and Steele, J.M. Defining and Measuring General Education Knowledge and Skills. Iowa City, IA: ACT, 1982.

Kitabchi, G. "Multivariate Analysis of Urban Community College Student Performance on the ACT College Outcomes Measures Program Test." Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1985. [ED#261-091]

Steele, J. M. "Assessing Speaking and Writing Proficiency via Samples of Behavior." Paper presented at the annual meeting of the Central States Speech Association, 1979. [ED#169-597]

Graduate Record Examinations Program: General Examinations

Publisher: Graduate Record Examinations Board, CN 6000, Princeton, NJ, 08541-6000; Scales: Verbal (antonyms, analogies, sentence completions, reading passages), Quantitative (quantitative comparisons, mathematics, data interpretation), Analytic (analytical reasoning, logical reasoning); Length: 50-76 items per sub-test; Time: 3 hours, 30 minutes.

The General Examinations of the GRE are nationally normed tests designed to assess learned abilities that are not related to any particular field of study, but that are related to the skills necessary for graduate study. Research on the GRE General Examinations has revealed high levels of reliability (.89 to .92) for the three tests. Reliability estimates for the nine item-types are somewhat lower (.60 to .90). Research has also found that test (and item-type) scores are related to undergraduate performance as well as to performance in graduate school.

Adelman, C. The Standardized Test Scores of College Graduates, 1964-1982. Washington, D.C.: U.S. Government Printing Office, 1984.

Conrad, L., Trismen, D., and Miller, R. Graduate Record Examinations Technical Manual. Princeton: ETS, 1977.

Fortna, R.O. Annotated Bibliography of the Graduate Record Examinations. Princeton: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1980.

Graduate Record Examinations Board. GRE Guide to the Use of the Graduate Record Examinations Program. Princeton: ETS, 1987.

Swinton, S.S. and Powers, D.E. A Study of the Effects of Special Preparation on GRE Analytical Scores and Item Types. GRE Research Report GREB 78-2R. Princeton: ETS, 1982.

Wilson, K.M. The Relationship of GRE General Test Item-Type Part Scores to Undergraduate Grades. GRE Research Report 81-22P. Princeton: ETS, 1985.

Assessment of Basic Skills

Descriptive Tests of Language Skills

Publisher: Descriptive Tests of Language Skills, Educational Testing Service, Mail Drop 22E, Princeton, NJ 08541; Scales: Reading Comprehension, Logical Relationships, Vocabulary, Usage, and Sentence Structure; Length: 30-50 items/test; Time: 15-30 minutes/test.

The Descriptive Tests of Language Skills (DTLS) consist of five tests designed for the placement of students in college English classes. These tests may be used separately or in combination. Because of their low difficulty levels, the DTLS are most appropriate for identifying students in need of remediation. Research by ETS indicates that all five tests evidence acceptable reliability (from .82 to .89); that the DTLS are correlated with writing ability and other measures of academic ability, such as ACT scores; and that performance on the DTLS predicts college grade point average. Studies have not examined the appropriateness of the DTLS as instruments for evaluating program quality.

College Entrance Examination Board. Guide to the Use of the Descriptive Tests of Language Skills. Princeton: ETS, 1985.

Snowman, J., Leitner, D.W., Snyder, V. and Lockhart, L., "A Comparison of the Predictive Validities of Selected Academic Tests of the American College Test (ACT) Assessment Program and the Descriptive Tests of Language Skills for College FreshmEn in a Basic Skills Program." Educational and Psychological Measurement, vol. 40, (1980), pp. 1159-1166.

Snyder, V. and Elmore, P.B., "The Predictive Validity of the Descriptive Tests of Language Skills for Developmental Students Over a Four-Year College Program." Educational and Psychological Measurement, vol. 43, (1983), pp. 1113-1122.

Descriptive Tests of Mathematics Skills

Publisher: Descriptive Tests of Mathematics Skills, Educational Testing Service, Mail Drop 22E, Princeton, NJ 08541; Scales: Arithmetic Skills, Elementary Algebra Skills, Intermediate Algebra Skills, and Functions and Graphs; Length: 30-35 items/test; Time: 30 minutes/test.

The four tests in the Descriptive Tests of Mathematics Skills (DMTS), used separately or in combination, are designed to assess mathematics skills for placement purposes. Because of their low item difficulty levels, these tests are not appropriate for differentiating among students with high levels of math skills. Research has indicated that the DTMS examinations evidence acceptable reliability (.84 to .91) and that the DTMS are related to measures of academic ability and performance in introductory math courses, particularly remedial courses.

Bridgeman, B., "Comparative Validity of the College Board Scholastic Aptitude Test--Mathematics and the Descriptive Tests of Mathematics Skills for Predicting Performance in College Mathematics Courses." Educational and Psychological Measurement, vol. 42, (1982), pp. 361-366.

College Entrance Examination Board. Guide to the Use of the Descriptive tests of Mathematics Skills. Princeton: ETS, 1985.

New Jersey College Basic Skills Placement Tests

Publisher: NJCBSPT, College Entrance Examination Board, Educational Testing Service, Princeton, NJ 08541; Scales: Writing, Reading Comprehension, Sentence Sense, Math Computation, Elementary Algebra, Composition (composite score), and Total English (composite score); Length: 168 items; Time: 3 hours.

The New Jersey College Basic Skills Placement Test (NJCBSPT) consists of five tests designed to meet the requirements of the assessment and evaluation program developed by the New Jersey Board of Higher Education. In addition to the five test scores, two composite scores can be derived from the language assessment parts of the test. Reliability estimates for the seven subscales range from .83 to .92. The content validity of the NJCBSPT was achieved by providing for constant review during test construction by a panel of experts from the New Jersey Basic Skills Council. Studies on the construct validity and predictive validity of the NJCBSPT are currently underway.

College Entrance Examination Board. The New Jersey College Basic Skills Placement Test Program: Your Information Base for Outcomes Assessment. Princeton: ETS, 1987.

Office of College Outcomes. Appendices to the Report of the New Jersey Board of Higher Education from the Advisory Committee to the College Outcomes Evaluation Program. Trenton, NJ: New Jersey Department of Higher Education, 1987.

Test of Standard Written English

Publisher: Test of Standard Written English, College Entrance Examination Board, Princeton, NJ 08541; Scales: Total Score; Length: 50 items; Time: 30 minutes.

The Test of Standard Written English (TSWE) is designed to measure a student's ability to use the language contained in most college textbooks. Research has found that the TSWE evidences acceptable reliability, and is predictive of performance in freshman English courses. The TSWE also has been found to be predictive of performance during the Junior year. Indeed, the TSWE has been found to be as good a predictor of performance as longer, more complex exams.

Bailey, R.L. "The Test of Standard Written English: Another Look." Measurement and Evaluation in Guidance, vol 10, (1977), pp. 70-74.

Michael, W.B. and Shaffer, P. "A Comparison of the Validity of the Test of Standard Written English (TSWE) and of the California State University and Colleges English Placement Test (CSUC-EPT) in the Prediction of Grades in a Basic English Composition Course and of Overall Freshman-Year Grade Point Average." Educational and Psychological Measurement, vol. 39, (1979), pp. 131-145.

Suddick, D.E. "A Re-examination of the Use of the Test of Standard Written English and Resulting Placement for Older Upper-Division and Master's Level Students." Educational and Psychological Measurement, vol. 42, (1982), pp. 367-369.

Suddick, D.E., "The Test of Standard Written English and Resulting Placement Patterns: A Follow-up of Performance of Older Upper-Division and Master Level Students." Educational and Psychological Measurement, vol. 41, (1981), pp 599-601.

Assessment of Cognitive Development

Analysis of Argument

Author: David G. Winter, Department of Psychology, Wesleyan University, Middletown, CT 06457; Scales: Total Score; Length: two exercises; Time: 10 minutes.

The Analysis of Argument is a production measure designed to assess clarity and flexibility of thinking skills. After reading a passage representing a particular position on a controversial issue, subjects are asked to write a response disagreeing with the original position. After 5 minutes, they are then instructed to write a short essay that agrees with the original position. The two essays are scored using a 10-category scheme. Because inter-rater agreement is a function of training, the authors do not provide estimates of reliability. The authors do report that studies have found that scores on the Analysis of Argument test are significantly related to other measures of cognitive development, as well as to previous educational experiences.

Stewart, A.J. and Winter, D.G. Analysis of Argument: An Empirically Derived Measure of Intellectual Flexibility. Boston: McBer and Company, 1977.

Erwin Scale of Intellectual Development

Author: T. Dary Erwin, Office of Student Assessment, James Madison University, Harrisonburg, VA 22801; Scales: Dualism, Relativism, Commitment, Empathy; Length: 86 items; Time: untimed.

The Erwin Scale of Intellectual Development (SID) was designed to measure intellectual development based on Perry's scheme, three of the four sub-scales (dualism, relativism and commitment) paralleling Perry's categories of intellectual development. Research on the SID has found that all four sub-scales evidence acceptable reliability (.70 to .81) and that the SID is significantly related to other measures of development, including measures of identity and involvement.

Erwin, T.D., "The Scale of Intellectual Development: Measuring Perry's Scheme." Journal of College Student Personnel, vol. 24, (1983), pp. 6-12.

Perry, W.G., Jr. Forms of Intellectual and Ethical Development in the College Years. New York: Holt, Rinehart and Winston, 1970.

Measure of Epistemological Reflection

Author: Margaret Baxter-Magolda, Department of Educational Leadership, Miami University, Miami, OH; Scales: Total Score; Length: 6 stimuli; Time: untimed.

The Measure of Epistemological Reflection (MER) represents a bridge between recognition and production measures. Six stimuli corresponding to Perry's levels of development are presented to subjects, who are then asked to justify the reasoning used in each stimulus. Standardized scoring procedures provide a quantified measure of intellectual development. Alpha reliability for the ratings may be as high as .76, while interrater reliability has ranged from .67 to .80, depending on the amount of training provided to raters. Research has provided support for the developmental underpinnings of the MER, revealing significant score differences for different educational levels.

Baxter-Magolda, M. and Porterfield, W.D. "A New Approach to Assess Intellectual Development on the Perry Scheme." Journal of College Student Personnel, vol. 26, (1985), pp. 343-351.

Reflective Judgment Interview

Authors: K.S. Kitchener, School of Education, University of Denver, Denver, CO, and P.M. King, Department of College Student Personnel, Bowling Green State University, Bowling Green, OH; Scales: Total Score; Length: four dilemmas; Time: approximately 40 minutes.

Like the MER, the Reflective Judgment Interview (RJI) represents a bridge between recognition and production measures. It consists of four dilemmas which are presented individually to the subject. Each dilemma is followed by a series of standardized questions designed to identify which of Perry's seven stages of intellectual development is being used by the subject to deal with that dilemma. A subject's score is the average rating across dilemmas and across raters. Research has shown that the RJI evidences acceptable levels of reliability (.73 to .78). In addition, the RJI has been found to be significantly related to other measures of critical thinking, as well as to levels of education.

Brabeck, M.M. "Critical Thinking Skills and Reflective Judgement Development: Redefining the Aims of Higher Education." *Journal of Applied Developmental Psychology*, vol. 4, (1983), pp. 23-34.

King, P.M. and Kitchener, K.S. "Reflective Judgment Theory and Research: Insights into the Process of Knowing in the College Years." Paper presented at the annual meeting of the American College Personnel Association, Boston, 1985. [ED#263-821]

Kitchener, K.S., and King, P.M. "Reflective Judgment: Concepts of Justification and Their Relationship to Age and Education." *Journal of Applied Developmental Psychology*, vol. 2, (1981), pp. 89-116.

Test of Thematic Analysis

Author: David G. Winter, Department of Psychology, Wesleyan University, Middletown, CT 06457; *Scales*: Total Score (optional: differentiation, discrimination, integration); *Length*: one exercise; *Time*: approximately 30 minutes.

The Test of Thematic Analysis uses a compare and contrast format to assess critical thinking skills. Subjects are presented with two sets of data and are asked to describe (in writing) how the two sets differ. The content of the essays is scored on a nine-point scale. In addition, scales derived from human information processing research can be used to evaluate the structure of the responses. Studies have found high levels of interrater agreement when scoring the TTA. Test scores also have been found to be significantly correlated with academic ability and coursework. In addition, measures of the structural characteristics of students' essays have been found to be significantly related to other measures of critical thinking, as well as to previous educational experiences.

Schroder, H.M., Driver, M.J. and Streufert, S. *Human Information Processing*. New York: Holt, Rinehart and Winston, 1967.

Winter, D.G. *Thematic Analysis: An Empirically Derived Measure of Critical Thinking*. Boston: McBer and Company, 1967.

Winter, D.G. and McClelland, D.C. "Thematic Analysis: An Empirically Derived Measure of the Effects of Liberal Arts Education." Journal of Educational Psychology, vol. 70, (1978), pp. 8-16.

Winter, D.G., McClelland, D.C. and Stewart, A.J. A New Case for the Liberal Arts. San Francisco: Jossey-Bass, 1981.

Watson-Glaser Critical Thinking Appraisal

Publisher: G. Watson and E.M. Glaser, Harcourt, Brace, and World; New York, NY; Scales: Total Score, Inference Recognition of Assumptions, Deduction, Interpretation, and Evaluation of Arguments; Length: 100 items; Time: 50 minutes.

The Watson-Glaser Critical Thinking Appraisal (CTA) is a multiple-choice measure designed to assess students' critical thinking abilities. In addition to a total score, five sub-scores can be derived from the CTA. Research has found that the total score on the CTA evidences acceptable reliability (.85 to .87) over seven norm groups and that students' performance on the CTA is positively related to their college experiences. In addition, the CTA has been found to be predictive of performance in courses emphasizing critical thinking.

Crites, J.O. "Test Review." Journal of Counseling Psychology, vol. 12, (1965), pp. 328-330.

Helmstadter, G.C. "Watson-Glaser Critical Thinking Appraisal." Journal of Educational Measurement, vol. 2, (1965), pp. 254-256.

Westbrook, B.W. and Sellers, J.R. "Critical Thinking, Intelligence, and Vocabulary." Educational and Psychological Measurement, vol. 27, (1967), pp. 443-446.

Wilson, D.G. and Wagner, E.E. "The Watson-Glaser Critical Thinking Appraisal as a Predictor of Performance in a Critical Thinking Course." Educational and Psychological Measurement, vol. 41, (1981), pp. 1319-1322.

Assessment of Values

Defining Issues Test

Author: James R. Rest, Department of Social, Psychological and Philosophical Foundations of Education, 330 Burton Hall, University of Minnesota, Minneapolis, MN 55455; Scales: "p" score; Length: 72 items; Time: untimed.

Rest developed the Defining Issues Test (DIT), a recognition measure of moral reasoning, based on the six stages identified by Kohlberg. Research has indicated that alpha reliability for the DIT is .77 and test-retest reliability is approximately .80. Research also has indicated that the DIT is significantly correlated with other measures of moral development, specifically Kohlberg's measure, and longitudinal research has found evidence of progression from lower-ordered to principled reasoning. Results also indicate that the DIT produces higher scores for principled reasoning than does Kohlberg's measure, and these higher scores are not due to upward faking on the DIT. These results suggest that production and recognition measures provide significantly different views of moral reasoning.

Biggs, D.A. and Barnett, R. "Moral Judgement Development of College Students." *Research in Higher Education*, vol. 14, (1981), pp. 91-102.

Davison, M.L. and Robbins, S. "The Reliability and Validity of Objective Indices of Moral Development." *Applied Psychological Measurement*, vol. 2, (1978), pp. 391-403.

McGeorge, C. "Susceptibility to Faking the Defining Issues Test of Moral Development." *Developmental Psychology*, vol. 11, (1975), p. 108.

Rest, J.R. "Longitudinal Study of the Defining Issues Test of Moral Judgement: A Strategy for Analyzing Developmental Change." *Developmental Psychology*, vol. 11, (1975), pp. 738-748.

Rest, J.R. *Development in Judging Moral Issues*. Minneapolis: University of Minnesota Press, 1979.

Rest, J.R., Cooper, D., Coder, R., Massanz, J. and Anderson, D. "Judging the Important Issues in Moral Dilemmas--An Objective Measure of Development." *Developmental Psychology*, vol. 10, (1974), pp. 491-501.

Humanitarian/Civic Involvement Values

Author: Ernest T. Pascarella, College of Education, University of Illinois at Chicago, Box 4348, Chicago, IL 60680; Scales: Total Score; Length: 6 items; Time: untimed.

The measure was derived from questions on the survey designed by the Cooperative Institutional Research Program (CIRP). Alpha reliability for this scale has been estimated to be .77. Results of research using this scale indicate that collegiate academic and social experiences are significantly related to the development of humanitarian/civic-involvement values, and that social involvement has the greater impact.

Pascarella, E.T., Ethington, C.A. and Smart, J.C. "The Influence of College on Humanitarian/Civic-Involvement Values." Paper presented at the annual meeting of the American Edcational Research Association, Washington, D.C., 1987.

Kohlberg's Measure of Moral Development

Author: Lawrence Kohlberg, "The Development of Modes of Moral Thinking and Choice in the Years Ten to Sixteen." Unpublished Doctoral Dissertation, University of Chicago, 1958; *Scales*: Total Score; *Length*: three dilemmas; *Time*: untimed.

In an effort to assess moral reasoning, Kohlberg developed a production measure that presents subjects with three moral dilemmas and requires them to explain how the dilemmas should be resolved. Subjects' responses are scored by raters trained to identify the dominant stage of moral reasoning employed. Reliability estimates for this technique are well within accepted limits (above .90). Research has provided support for the construct validity of Kohlberg's approach, identifying a clear step-by-step progression through the stages of moral reasoning. Moral reasoning also has been linked to students' previous educational experiences.

Kohlberg, L. *The Psychology of Moral Development*. New York: Harper and Row, 1984.

Rokeach Value Survey

Publisher: Halgren Tests, The Free Press, New York, NY; *Scales*: Instrumental Values, Terminal Values; *Length*: 36 items; untimed.

The Rokeach Value Survey was designed as a means of describing subjects' value systems. Respondents are asked to rank two sets of values (instrumental and terminal). Multiple administrations of the instrument can be used to measure stability and change in value systems. Test-retest reliability has been estimated to be adequate (.65 to .74). Moreover, research has shown that changes in individuals' value systems can be linked to life events.

Rokeach, M. *The Nature of Human Values*. New York: The Free Press, 1973.

Rokeach, M. (ed.) *Understanding Human Values: Individual and Societal*. New York: The Free Press, 1979.

☆ U.S. GOVERNMENT PRINTING OFFICE: 1988— 2 1 7 -0 4 4 /0

United States
Department of Education
Washington, DC 20208

Official Business
Penalty for Private Use, $300

Postage and Fees Paid
U.S. Department of Education
Permit No. G-17

Third Class

OR 88-514